Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin

Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling

Rémi Gribonval* remi.gribonval@inria.fr
Gilles Blanchard † gilles.blanchard@universite-paris-saclay.fr
Nicolas Keriven‡ nicolas.keriven@gipsa-lab.grenoble-inp.fr
Yann Traonmilin§ yann.traonmilin@u-bordeaux.fr

June 11, 2021

Abstract

We provide statistical learning guarantees for two unsupervised learning tasks in the context of *compressive statistical learning*, a general framework for resource-efficient large-scale learning that we introduced in a companion paper. The principle of compressive statistical learning is to compress a training collection, in one pass, into a low-dimensional *sketch* (a vector of random empirical generalized moments) that captures the information relevant to the considered learning task. We explicitly describe and analyze random feature functions which empirical averages preserve the needed information for *compressive clustering* and *compressive Gaussian mixture modeling* with fixed known variance, and establish sufficient sketch sizes given the problem dimensions.

Keywords: Kernel mean embedding, random features, random moments, statistical learning, dimension reduction, unsupervised learning, clustering, mixture modeling, principal component analysis.

1 Introduction

Motivated by the need to handle large-scale learning in streaming or distributed contexts with limited memory, we studied in a companion paper [Gribonval et al., 2021] a general *compressive learning framework* based on a generic sketching mechanism, using empirical averages of a random feature function to compress a whole training collection into a single *sketch* vector. Learning from such a sketch is expressed as (generalized) moment fitting problem. Statistical learning guarantees control the excess risk of the overall procedure, provided the used random feature function satisfies certain properties with respect to the considered learning task. The size of the sketch can also be controlled.

For compressive clustering, aka compressive k-means [Keriven et al., 2017], and compressive Gaussian mixture modeling [Keriven et al., 2016, 2018], good empirical results have been obtained with compressive statistical learning using *random Fourier moments*, i.e. using empirical averages of random Fourier features [Rahimi and Recht, 2008]. Based on the general framework of [Gribonval et al.,

*Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France
This work was initiated while R. Gribonval, N. Keriven and Y. Traonmilin were with Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes, France
†Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay F-91405, Orsay, France.
‡CNRS, GIPSA-lab, UMR 5216, F-38400 Saint-Martin-d'Hères, France
§CNRS, Univ. Bordeaux, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France.

2021], we establish that these empirical results are supported by statistical learning guarantees controlling the excess risk and the sketch sizes as a function of the problem dimensions.

For **compressive clustering** in dimension d, we demonstrate that a sketch of size

$$m \geq Ck^2 d \cdot \log^2 k \cdot (\log(kd) + \log(R/\varepsilon)),$$

with k the prescribed number of clusters, R a bound on the norm of the centroids, ε the separation between them, and C some universal constant, is sufficient to obtain statistical guarantees. To the best of our knowledge these are the first guarantees of this kind: while there is a substantial literature on asymptotical and nonasympotical guarantees on convergence rates for clustering Pollard [1982b], Chou [1994], Linder et al. [1994], Bartlett et al. [1998], Antos et al. [2005], Antos [2005], Fischer [2010], Levrard [2013], they are based on the full data while our focus is on analysis of sketching.

For **compressive Gaussian mixture estimation** with known covariance in dimension d, we identify a finite sketch size sufficient to obtain statistical guarantees under a separation assumption between means expressed in the Mahalanobis norm associated to the known covariance matrix. A parameter embodies the tradeoff between sketch size and separation. At one end of the spectrum the sketch size is quadratic in k and exponential in d and guarantees are given for means that can be separated in $\mathcal{O}\big(\sqrt{\log k}\big)$. This compares favorably to existing literature [Achlioptas and McSherry, 2005, Vempala and Wang, 2004] (recent works make use of more complex conditions that theoretically permits arbitrary separation [Belkin and Sinha, 2010], however all these approaches use the full data while we consider a compressive approach that uses only a sketch of the data). At the other end the sketch size is quadratic in k and *linear* in d, however the required separation is of the order of $\sqrt{d \log k}$.

After recalling the outline of the general framework for compressive statistical learning of [Gribonval et al., 2021] in Section 2, a sketching procedure and the associated learning guarantees for compressive clustering (respectively for compressive Gaussian mixture estimation) is given in Section 3 (respectively Section 4). The rest of the paper is dedicated to establishing these results. Section 5 describes a generic approach to establish learning guarantees when estimation of mixtures of elementary distributions are involved, as in the two considered examples. The results are then specialized in Section 6 to mixtures based on location families, using weighted random Fourier features. The most technical proofs are postponed to appendices, including the proof of the main results of Sections 3 and 4.

2 Overview of compressive statistical learning

In statistical learning, one is given a training collection $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathcal{Z}^n$ assumed to be drawn i.i.d. from a probability distribution π on the measurable space $(\mathcal{Z}, \mathfrak{Z})$. In our examples, $\mathcal{Z} = \mathbb{R}^d$ endowed with the Borel sigma-algebra $\mathfrak{Z}$. A learning task (supervised or unsupervised) is formally defined through a *loss function* $\ell : (x, h) \mapsto \ell(x, h) \in \mathbb{R}$ which measures how adapted is a training sample x to a hypothesis h from some hypothesis class $\mathcal{H}$. The overall goal is to select a hypothesis $h^\star$ minimizing the *expected risk* $\mathcal{R}(\pi, h) := \mathbb{E}_{X\sim\pi}\, \ell(X, h)$,

$$h^\star \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h). \tag{1}$$

Remark 2.1. *We will always implicitly restrict our attention to probability distributions π that are $\mathcal{L}(\mathcal{H})$-integrable, i.e. such that $x \mapsto \ell(x, h)$ is measurable and π-integrable for all $h \in \mathcal{H}$.*

In practice, since the expected risk cannot be computed from the training collection, a common strategy is instead to minimize the *empirical risk* $\mathcal{R}(\hat{\pi}_n, h)$ (or a regularized version) associated to the *empirical probability distribution* $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ of the training samples. The two primary tasks considered in this paper are:

- **k-means (resp. k-medians) clustering**: each hypothesis h corresponds to a set of (at most) k candidate cluster centers, $c_1, \dots, c_k$, and the loss is defined by the k-means cost $\ell(x, h) = \min_{1 \leq l \leq k} \|x - c_l\|_2^2$, (resp. the k-medians cost $\ell(x, h) = \min_{1 \leq l \leq k} \|x - c_l\|_2$). The hypothesis class $\mathcal{H}$ may be further reduced by defining constraints on the considered centers (e.g., in some domain, or as we will see with some separation between centers).
- **Gaussian Mixture Modeling with fixed covariance matrix $\mathbf{\Sigma}$**: each hypothesis h corresponds to the collection of weights α_ℓ and means c_ℓ of a mixture of k Gaussians $\pi_{c_\ell} := \mathcal{N}(c_\ell, \mathbf{\Sigma})$, which probability density function is denoted $\pi_h(x) = \sum_{\ell=1}^k \alpha_\ell \mathcal{N}(c_\ell, \mathbf{\Sigma})$. The mixture parameters may again further be constrained by boundedness or separation assumptions. The loss function $\ell(x, h) = -\log \pi_h(x)$ is associated to the maximum likelihood estimation principle.

2.1 Principles of *compressive* statistical learning

Compressive learning proposes instead to choose a measurable (nonlinear) *feature function* $\Phi : \mathcal{Z} \mapsto \mathbb{R}^m$ or $\mathbb{C}^m$ and to proceed in two steps:

1. Compute empirical averages of the feature function to obtain a *sketch vector*

$$\mathbf{y} := \texttt{Sketch}(\mathbf{X}) := \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \in \mathbb{R}^m \text{or } \mathbb{C}^m; \tag{2}$$

2. Produce a hypothesis from the sketch by solving an optimization problem

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} R(\mathbf{y}, h) \tag{3}$$

with some adequate proxy $R(\mathbf{y}, \cdot)$ for the empirical risk $\mathcal{R}(\hat{\pi}_n, \cdot)$.

For the learning tasks considered in this paper, the feature function Φ is built using random Fourier features [Rahimi and Recht, 2008, 2009]. For compressive k-medians/k-means, the proxy is

$$R_{\texttt{clust.}}(\mathbf{y}, h) := \min_{\alpha \in \mathbb{S}_{k-1}} \left\| \sum_{\ell=1}^k \alpha_\ell \Phi(c_\ell) - \mathbf{y} \right\|_2, \tag{4}$$

with $\mathbb{S}_{k-1} := \left\{ \alpha \in \mathbb{R}^k : \alpha_\ell \geq 0; \sum_{\ell=1}^k \alpha_\ell = 1 \right\}$ the simplex. For compressive GMM, it reads

$$R_{\texttt{GMM}}(\mathbf{y}, h) := \left\| \sum_{\ell=1}^k \alpha_\ell \Psi(c_\ell) - \mathbf{y} \right\|_2. \tag{5}$$

with $\Psi(c_\ell) := \mathbb{E}_{X \sim \mathcal{N}(c_\ell, \mathbf{\Sigma})} \Phi(X)$. The nonlinear parametric optimization problems corresponding to the minimization of (4)-(5) have be empirically addressed with success using continuous analogs of greedy algorithms for sparse reconstruction in inverse linear problems [Keriven et al., 2016, 2017, 2018].

2.2 Statistical learning guarantees for compressive statistical learning

From a statistical learning perspective, the goal is to control the *excess risk* $\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^\star)$. In compressive statistical learning, this requires measuring a "distance" between the data distribution π and some *model set* $\mathfrak{S}$. Specific instances of model sets considered in this paper are:

- for compressive k-means / k-medians: mixtures of k (separated) Diracs;

- for compressive Gaussian mixture modeling: mixtures of k (separated) Gaussians.

An important conceptual tool is the so-called *sketching operator* $\mathcal{A}$ defined by

$$\mathcal{A}(\pi) := \mathbb{E}_{X\sim\pi}\Phi(X) \tag{6}$$

which is linear in the sense that[1] $\mathcal{A}(\theta\pi+(1-\theta)\pi') = \theta\mathcal{A}(\pi)+(1-\theta)\mathcal{A}(\pi')$ for any π, π' and $0 \leq \theta \leq 1$. A key property of this operator for compressive statistical learning is the preservation of certain task-driven metrics on probability distributions from the considered model set $\mathfrak{S}$. Denoting

$$\Delta\mathcal{R}_{h_0}(\pi, h) := \mathcal{R}(\pi, h) - \mathcal{R}(\pi, h_0), \tag{7}$$

the excess risk relative to a reference hypothesis h_0, the *excess risk divergence* with respect to h_0 is

$$D^{\mathcal{H}}_{h_0}(\pi\|\pi') := \sup_{h\in\mathcal{H}}(\Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(\pi', h)) \geq (\Delta\mathcal{R}_{h_0}(\pi, h_0) - \Delta\mathcal{R}_{h_0}(\pi', h_0)) = 0. \tag{8}$$

Given a class $\mathcal{G}$ of measurable functions $g : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$, we denote

$$\|\pi - \pi'\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}}|\mathbb{E}_{X\sim\pi}g(X) - \mathbb{E}_{X'\sim\pi'}g(X')|. \tag{9}$$

Specializing this to the particular class $\mathcal{G} = \Delta\mathcal{L}(\mathcal{H})$, where

$$\Delta\mathcal{L}(\mathcal{H}) := \{\ell(\cdot, h) - \ell(\cdot, h') : h, h' \in \mathcal{H}\}, \tag{10}$$

we get a task-driven metric $\|\pi - \pi'\|_{\Delta\mathcal{L}(\mathcal{H})} = \sup_{h_0} D^{\mathcal{H}}_{h_0}(\pi\|\pi')$, capturing differences in terms of excess risks. The first main result of [Gribonval et al., 2021] that we will use is the following theorem.

Theorem 2.2 ([Gribonval et al., 2021, Theorem 2.5])**.** *Consider a loss class $\mathcal{L}(\mathcal{H}) := \{\ell(\cdot, h) : h \in \mathcal{H}\}$, a feature function Φ, and a model set $\mathfrak{S}$ that is both $\mathcal{L}(\mathcal{H})$-integrable and $\{\Phi\}$-integrable (cf Remark 2.1). Assume that the sketching operator $\mathcal{A}$ associated to Φ satisfies the following* lower restricted isometry property (LRIP)

$$\|\tau' - \tau\|_{\Delta\mathcal{L}(\mathcal{H})} \leq C_{\mathcal{A}}\|\mathcal{A}(\tau') - \mathcal{A}(\tau)\|_2 + \eta \quad \forall\tau, \tau' \in \mathfrak{S} \tag{11}$$

for some finite positive constants $C_{\mathcal{A}}$ and η.

Consider any training collection $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathcal{Z}^n$ and denote $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Define

$$\mathbf{y} := \texttt{Sketch}(\mathbf{X}) = \mathcal{A}(\hat{\pi}_n), \tag{12}$$

$$\widetilde{\pi} \in \mathfrak{S} \text{ satisfying } \|\mathcal{A}(\widetilde{\pi}) - \mathbf{y}\|_2 \leq (1+\nu)\inf_{\tau\in\mathfrak{S}}\|\mathcal{A}(\tau) - \mathbf{y}\|_2 + \nu', \quad \text{for some constants } \nu, \nu' \geq 0, \tag{13}$$

$$\hat{h} \text{ satisfying } \mathcal{R}(\widetilde{\pi}, \hat{h}) \leq \inf_{h\in\mathcal{H}}\mathcal{R}(\widetilde{\pi}, h) + \varepsilon', \quad \text{for some constant } \varepsilon' \geq 0. \tag{14}$$

Then, for any probability distribution π that is both $\mathcal{L}(\mathcal{H})$-integrable and $\{\Phi\}$-integrable:

$$\forall h_0 \in \mathcal{H} : \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) \leq d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S}) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 + \eta + C_{\mathcal{A}}\nu' + \varepsilon', \tag{15}$$

where

$$d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S}) := \inf_{\tau\in\mathfrak{S}}\big(D^{\mathcal{H}}_{h_0}(\pi\|\tau) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2\big). \tag{16}$$

[1] One can indeed extend $\mathcal{A}$ to a linear operator on the space of finite signed measures such that Φ is integrable, see [Gribonval et al., 2021, Appendix A.2].

The bound (15) holds regardless of any distribution assumption on the training collection $\mathbf{X}$, and is valid for any $h_0 \in \mathcal{H}$. The estimate is of course primarily of interest when $\mathbf{X}$ is drawn i.i.d. from π and with $h_0 = h^\star$, in which case the left-hand side is the excess risk with respect to the optimum hypothesis, and $\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2$ typically decays as $1/\sqrt{n}$. As we will see, other choices of h_0 will nevertheless turn out to be useful for the analysis of compressive k-means and compressive Gaussian Mixture Modeling, where we need to introduce a class $\mathcal{H}$ satisfying a separation assumption and may be interested in the best hypothesis over a larger hypothesis class $\bar{\mathcal{H}} \supseteq \mathcal{H}$.

The model sets $\mathfrak{S}$ considered for compressive k-means/k-medians consist of the probability distributions for which the optimum risk vanishes, i.e. $\mathcal{R}(\pi, h^\star) = 0$. These model sets are precisely chosen to ensure that the bias term (16) in the control (15) of the excess risk vanishes when the optimum risk itself vanishes, and we will provide more explicit bounds on this bias term. With these model sets, solving (13) and (14) with $\nu = \nu' = \varepsilon' = 0$ also precisely corresponds to minimizing the proxy (4). For compressive GMM, the considered model set $\mathfrak{S}$ consists of mixtures of Gaussians. Again, solving (13) and (14) with $\nu = \nu' = \varepsilon' = 0$ corresponds to minimizing the proxy (5).

The main technical contribution of this paper is to establish that the main assumption (11) of Theorem 2.2 holds for compressive clustering (resp. compressive GMM), using sketching operators based on random Fourier features with controlled sketch size m. For this, we rely on the general approach described in [Gribonval et al., 2021, Section 5] relating random (Fourier) features and kernel mean embeddings of probability distributions. Another contribution is to provide more concrete estimates of the "bias term" (16). The results are first stated in the next sections, before giving the technical ingredients for their proof in the rest of the paper.

3 Compressive Clustering

We consider here two losses that measure clustering performance: the **k-means** and **k-medians** losses. Hypotheses are k-tuples $(c_1, \ldots, c_k)$ where the elements $c_l \in \mathbb{R}^d$ are the so-called centers of clusters. We speak of *unconstrained* k-means or k-medians if the cluster centers can be arbitrary points of $\mathbb{R}^{d\times k}$, and *constrained* otherwise (if the k-tuple of centers must belong to a specific subset of $\mathbb{R}^{d\times k}$, for instance if there is a separation or a maximum norm constraint). The loss function for the clustering task is

$$\ell(x, h) := \min_{1\leq l\leq k} \|x - c_l\|_2^p \tag{17}$$

with $p = 2$ for k-means (resp. $p = 1$ for k-medians) and $\mathcal{R}(\pi, h) = \mathbb{E}_{X\sim\pi} \min_{1\leq l\leq k} \|X - c_l\|_2^p$.

3.1 Main theoretical guarantees

Existence and properties of a minimizer of the risk. For unconstrained k-means clustering, given any $\mathcal{L}(\mathcal{H})$-integrable probability distribution π there exists a global minimizer h_π^* of the risk, see e.g [Pollard, 1982a, Lemma 8] which only assumes that the support of the distribution π contains at least k elements. When this support has at most $k-1$ elements the existence of a global minimizer is trivial, and a zero risk is achieved. The existence of a global optimum was also proved in more general settings that include unconstrained k-medians clustering, see, e.g. [Graf et al., 2007, Theorem 1]. For unconstrained k-means, when the support of π contains at least k elements, the global minimizer satisfies necessary conditions that were already identified in the work of Steinhaus [1956] and were formalized more recently in a generalized setting [Graf et al., 2007, Proposition 1]. For a generic hypothesis $h = (c_1, \ldots, c_k)$ denote

$$V_l(h) := \left\{ x \in \mathbb{R}^d : \|x - c_l\|_2 = \min_j \|x - c_j\|_2 \right\}, \qquad 1 \leq l \leq k \tag{18}$$

the Voronoi cells of the l-th center. Let $(W_j(h))_{1\leq h\leq k}$ be an arbitrary *Voronoi partition* associated to these Voronoi cells, i.e., $W_j(h) \subseteq V_j(h)$ and $(W_j(h))_{1\leq j\leq k}$ form a partition of $\mathbb{R}^d$ (in other words, this partition breaks the "ties" at the boundary of the Voronoi cells arbitrarily). For $x \in \mathcal{Z} = \mathbb{R}^d$, let $P_h x = c_j$ if and only if $x \in W_j(h)$ (i.e. P_h maps x to the closest cluster center with tie-breaking given by the Voronoi partition), and $P_h\pi$ be the push-forward of a probability distribution π through P_h. In other words, putting $\alpha_l(\pi, h) := \pi(X \in W_l(h)) = \mathbb{E}_{X\sim\pi}\mathbf{1}_{W_l(h)}(X)$ the probability that a sample belongs to a piece of the Voronoi partition, with $\mathbf{1}_E$ the indicator function of set E, we have $P_h\pi = \sum_{i=1}^k \alpha_i(\pi, h)\delta_{c_i}$.

For unconstrained k-means and π with a support containing at least k points, the k optimal centers associated to $h^\star_\pi$ are pairwise distinct ($c_i \neq c_j$ for $i \neq j$) and satisfy the so-called centroid condition: for $1 \leq l \leq k$ we have $\alpha_l(\pi, h^\star) > 0$ and

$$c_l = \mathbb{E}_{X\sim\pi}(X|X \in W_l(h^\star)) = \frac{\mathbb{E}_{X\sim\pi}\mathbf{1}_{W_l(h^\star)}(X) \cdot X}{\alpha_l(\pi, h^\star)}. \tag{19}$$

Finally, the optimal centers are such that $\pi(X \in V_i(h^\star) \cap V_j(h^\star)) = 0$ for each $i \neq j$, i.e., the distinction between Voronoi cells and partition pieces becomes essentially moot at the optimum.

Choice of a model set. Both k-means and k-medians are "compression-type" tasks [Gribonval et al., 2021, Definition 3.1] (see reminders in Appendix D.3). Given a hypothesis $h = (c_1, \ldots, c_k)$, the distributions such that $\mathcal{R}(\pi, h) = 0$ are precisely mixtures of k Diracs,

$$\mathfrak{S}^{\mathtt{CT}}_h = \left\{ \sum_{l=1}^k \alpha_l \delta_{c_l} : \alpha \in \mathbb{S}_{k-1} \right\} \tag{20}$$

where $\mathbb{S}_{k-1} := \left\{\alpha \in \mathbb{R}^k : \ \alpha_l \geq 0, \sum_{l=1}^k \alpha_l = 1\right\}$ denotes the $(k-1)$-dimensional simplex. Given a hypothesis class $\mathcal{H} \subseteq (\mathbb{R}^d)^k$, following the approach outlined in [Gribonval et al., 2021, Section 3.2], we consider the model set $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H}) := \cup_{h\in\mathcal{H}}\mathfrak{S}^{\mathtt{CT}}_h$.

Choice of a sketching function. Given that each model set $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$ consists of mixtures of Diracs, and by analogy with compressive sensing where random Fourier sensing yields RIP guarantees, it was proposed [Keriven et al., 2017] to perform compressive clustering using random Fourier moments. To establish our theoretical guarantees we rely on a reweighted version where the feature function $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ is defined as:

$$\Phi(x) := \frac{1}{\sqrt{m}}\left[\frac{e^{\jmath\omega_j^T x}}{w(\omega_j)}\right]_{j=1,\ldots,m} \quad \text{with } w(\omega) := 1 + \frac{s^2\|\omega\|_2^2}{d} \tag{21}$$

where $s > 0$ is a scale parameter. The random frequencies $\omega_1, \ldots, \omega_m$ in $\mathbb{R}^d$ are sampled independently from the distribution with density

$$\Lambda(\omega) = \Lambda_{w,s}(\omega) \propto w^2(\omega)e^{-\frac{s^2\|\omega\|^2}{2}}, \tag{22}$$

This sketching operator is based on a reweighting of Random Fourier Features [Rahimi and Recht, 2008]. The weights $w(\omega)$ are required for technical reasons (see general proof strategy in Section 5) but may be an artefact of our proof technique.

Learning from a sketch by minimizing a proxy for the risk. For any distribution in the model set, $\widetilde{\pi} = \sum_{l=1}^{k} \alpha_l \delta_{c_l} \in \mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$, the optimum of minimization (14) (with $\varepsilon' = 0$) is achieved with $\hat{h} = (c_1, \ldots, c_k)$ hence, given a sketch vector $\mathbf{y}$ and a class of hypotheses $\mathcal{H}$, finding near minimizers of (13) and (14) in Theorem 2.2 corresponds to finding a (near) minimizer $\hat{h} = (\hat{c}_1, \ldots, \hat{c}_k) \in \mathcal{H}$ of the following proxy for the risk

$$R_{\mathtt{clust.}}(\mathbf{y}, h) := \min_{\alpha \in \mathbb{S}_{k-1}} \left\| \mathbf{y} - \sum_{i=1}^{k} \alpha_i \Phi(c_i) \right\|_2 \tag{23}$$

in a *constrained* hypothesis class $\mathcal{H}$ that we now describe.

Separation constraint and approximate optimization. Because one can show (see Lemma 3.5) that it is necessary to establish an LRIP, we optimize the proxy $R(\mathbf{y}, h)$ on a restricted hypothesis class

$$\mathcal{H}_{k,2\varepsilon,R} := \left\{ (c_l)_{l=1}^{k} : \quad \min_{c_l \neq c_{l'}} \|c_l - c_{l'}\|_2 \geq 2\varepsilon, \quad \max_{l} \|c_l\|_2 \leq R \right\}. \tag{24}$$

There can be exact repetitions ($c_l = c_{l'}$ with $l \neq l'$), however distinct centers $c_l \neq c_{l'}$ must be separated. The parameters ε and R represent the resolution and extent at which we seek clusters in the data through the minimization of $R(\mathbf{y}, h)$. Observe that $\mathcal{H}_{k,2\varepsilon,R}$ is non-empty whenever $0 \leq \varepsilon \leq R$. When R/ε is close to one, its elements may nevertheless be forced to have repeated entries. The following result is proved in Appendix D.4. Besides leveraging Theorem 2.2 the proof exploits a generic approach developed in Section 5 to establish the LRIP on mixture models using Theorem 5.1, and its specialization to mixtures based on location families, which is developed in Section 6. We remind the reader that P_h is the projection onto the centroids of h, as described at the beginning of this section.

Theorem 3.1. *Consider Φ as in (21) where the ω_j are drawn according to (22) with scale $s > 0$. Given an integer $k \geq 1$ define $\varepsilon := 4s\sqrt{\log(ek)}$ and consider $R \geq \varepsilon$.*

1. *There is a universal constant $C > 0$ such that, for any $\zeta, \delta \in (0,1)$, if the sketch size satisfies* [2]

$$m \geq C\delta^{-2}\Big[k^2 d \cdot \big(1 + \log kd + \log \tfrac{R}{\varepsilon} + \log \tfrac{1}{\delta}\big) + k \log \tfrac{1}{\zeta}\Big] \cdot \log(ke) \min(\log(ek), d), \tag{25}$$

   with probability at least $1 - \zeta$ on the draw of $(\omega_j)_{j=1}^{m}$ the operator $\mathcal{A}$ induced by Φ satisfies

$$1 - \delta \leq \frac{\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} \leq 1 + \delta, \qquad \forall \tau, \tau' \in \mathfrak{S}^{\mathtt{CT}}(\mathcal{H}_{k,2\varepsilon,R}). \tag{26}$$

   with $\|\cdot\|_\kappa$ the mean map discrepancy [3] *(MMD) associated to kernel $\kappa(x, y) \propto \exp(-\|x - y\|_2^2 / s^2)$.*

2. *If (26) holds then:*
   - *The function Φ is L-Lipschitz with $L = \sqrt{1+\delta}/s$ with respect to Euclidean norms.*
   - *Consider any samples $x_i \in \mathbb{R}^d$, $1 \leq i \leq n$ (represented by the empirical distribution $\hat{\pi}_n$) and any probability distribution π on $\mathbb{R}^d$ that is both $\mathcal{L}(\mathcal{H})$-integrable and $\{\Phi\}$-integrable.*

[2] The sketch sizes from the introduction and [Gribonval et al., 2021, Table 1] involve $\log(\cdot)$ factors which have correct order of magnitude when their argument is large, but vanish when their argument is one. Factors $\log(e\cdot)$ do not have this issue.

[3] see (48) in Section 5.1 for details.

Consider a constrained class $\mathcal{H} \subseteq \mathcal{H}_{k,2\varepsilon,R} \subseteq \overline{\mathcal{H}} := (\mathbb{R}^d)^k$ and denote $\mathcal{R}(\cdot, h)$ the risk associated to k-medians (resp. k-means), $h^\star \in \arg\min_{h\in\overline{\mathcal{H}}} \mathcal{R}(\pi, h)$, $\pi^\star := P_{h^\star}\pi$. Consider $\hat{h} \in \mathcal{H}$ and $\nu, \nu' > 0$ such that

$$R_{\texttt{clust.}}(\mathbf{y}, \hat{h}) \leq (1+\nu) \inf_{h\in\mathcal{H}} R_{\texttt{clust.}}(\mathbf{y}, h) + \nu'. \tag{27}$$

with the proxy $R_{\texttt{clust.}}(\mathbf{y}, \cdot)$ defined in (23) and the sketch vector $\mathbf{y} := \frac{1}{n}\sum_{i=1}^n \Phi(x_i) = \mathcal{A}(\hat{\pi}_n)$. The excess risk of $\hat{h}$ with respect to $h^\star$ satisfies

$$\Delta\mathcal{R}_{h^\star}(\pi, \hat{h}) \leq (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 + d(\pi^\star, \mathcal{H}) + C_{\mathcal{A}}\nu' \tag{28}$$

where $C_{\mathcal{A}} \leq 56\sqrt{k/(1-\delta)}(2R)^p$ (with $p = 1$ for k-medians, $p = 2$ for k-means) and

$$d(\pi^\star, \mathcal{H}) := \inf_{\tau\in\mathfrak{S}^{\text{CT}}(\mathcal{H})} \left\{ \sup_{h\in\mathcal{H}} (\mathcal{R}(\pi^\star, h) - \mathcal{R}(\tau, h)) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 \right\}. \tag{29}$$

The first term in (28) is a measure of statistical error that can be easily controlled since $\|\Phi(x)\|_2 \leq 1$ by construction (21). By the vectorial Hoeffding's inequality [Pinelis, 1992], if $x_i, 1 \leq i \leq n$ are drawn i.i.d. with respect to π then with high probability it holds that $\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 \lesssim 1/\sqrt{n}$. The constants ν and ν' measure the numerical error in optimizing $R_{\texttt{clust.}}(\mathbf{y}, h)$ over $\mathcal{H}$.

The second term measures how close π is to $\pi^\star = P_{h^\star}\pi$, i.e., how "clusterable" is π. By [Gribonval et al., 2021, Lemma 3.4] and the Lipschitz property of Φ this is bounded by $L \cdot \mathcal{R}^{1/p}(\pi, h^\star)$. This bound seems however pessimistic and we leave to future work a possible sharpening for certain settings.

The third term $d(\pi^\star, \mathcal{H})$ measures of how far $h^\star$ (weighted by the coefficients α_i) is from belonging to $\mathcal{H}$. The Lipschitz property of Φ yields more explicit bounds that may deserve to be further sharpened.

Lemma 3.2. *With the notations of Theorem 3.1 consider a mixture of k Diracs $\pi^\star := \sum_{i=1}^k \alpha_i \delta_{c_i}$ with $c_i \in \mathbb{R}^d$ and $\boldsymbol{\alpha} \in \mathbb{S}_{k-1}$. If (26) holds then for the k-medians clustering task we have*

$$d_{\texttt{k-medians}}(\pi^\star, \mathcal{H}) \leq C \cdot \inf_{h\in\mathcal{H}} \mathcal{R}_{\texttt{k-medians}}(\pi^\star, h),$$

while for the k-means clustering task we have

$$d_{\texttt{k-means}}(\pi^\star, \mathcal{H}) \leq \inf_{h\in\mathcal{H}} \left\{ \mathcal{R}_{\texttt{k-means}}(\pi^\star, h) + 4CR \cdot \mathcal{R}_{\texttt{k-medians}}(\pi^\star, h) \right\},$$

with $C := 1 + (2+\nu)500\sqrt{k\log(ek)(1+\delta)/(1-\delta)}\frac{R}{\varepsilon}$.

The proof is in Appendix D.5. By Jensen's inequality we have $\mathcal{R}_{k-\texttt{medians}}(\pi, h) \leq \sqrt{\mathcal{R}_{k-\texttt{means}}(\pi, h)}$ hence $d_{\texttt{k-means}}(\pi^\star, \mathcal{H})$ can also be bounded in terms of the optimum k-means risk over $\mathcal{H}$.

Remark 3.3. *Just as Theorem 3.1, the above lemma is valid for an arbitrary class $\mathcal{H} \subseteq \mathcal{H}_{k,2\varepsilon,R}$, hence they can be exploited for instance when the $d \times k$ matrix of centroids $\mathbf{C} = [c_1, \ldots, c_k]$ is constrained to satisfy certain structural constraints (besides 2ε-separation and R-boundedness). For example, $\mathbf{C}$ could be required to be a product of sparse matrices to enable accelerated clustering [Giffon et al., 2021]. This is however left to future work.*

Remark 3.4. *Even if $\pi^\star \in \mathcal{H}$ and the third term vanishes, the second term is in general positive and does not vanish with $n \to \infty$, except if $\mathcal{R}(\pi, h^\star) = 0$, i.e. the source distribution π is itself exactly a mixture of k Diracs. This comes from the fact that the proposed method finds an optimal clustering*

for an implicitly reconstructed distribution which is of this form: if the source distribution is not of this form, this introduces an inherent bias. Hence, the result above does not recover consistency or convergence rates for the clustering risk available under broad conditions when using the full data (see Introduction for references on this topic). An interesting direction left for future work is to investigate if consistency could be established when considering a larger model set such as mixtures of r Diracs for $r > k$, and r depending on n, similarly to considerations on sketched PCA [Gribonval et al., 2021, discussion following Theorem 4.1]

3.2 Role of the separation assumption

It is natural to wonder whether the separation assumption is an artefact of our proof technique. The following result shows that it is in fact necessary to establish an LRIP for compressive k-means and compressive k-medians clustering, for any smooth sketching operator (in particular one based on Fourier features). The proof is in Appendix C.6.

Lemma 3.5. *Consider a loss associated to a clustering task: $\ell(x, h) := \min_l \|x - c_l\|_2^p$ where $h = (c_1, \dots, c_k)$, $k \geq 2$, with $p = 1$ (k-medians) or $p = 2$ (k-means). Let Φ be a sketching function of class $\mathcal{C}^2$ and $\mathcal{A}$ be the associated sketching operator. There is a constant $c_\Phi > 0$ such that for any $R > 0$ and any $0 < \varepsilon \leq R$, with $\mathcal{H} = \mathcal{H}_{k,\varepsilon,R}$ we have*

$$\sup_{\tau,\tau' \in \mathfrak{S}^{\mathrm{ct}}(\mathcal{H})} \frac{\|\tau - \tau'\|_{\Delta\mathcal{L}(\mathcal{H})}}{\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2} \geq c_\Phi R^{p-1}/\varepsilon.$$

In particular, the LRIP (11) cannot hold with $\eta = 0$ and a finite constant C on $\mathcal{H}_{k,0,R}$.

Although this shows that the separation ε is important in the derivation of learning guarantees, its role in the final bounds is less stringent than it may appear at first sight, for several reasons.

First, in the most favorable case, the globally optimal hypothesis $h^\star$ indeed belongs to the constrained class $\mathcal{H}$, in which case $d(\pi^\star, \mathcal{H}) = 0$ since $\pi^\star = P_{h^\star}\pi$ is a mixture of Diracs. In particular, for $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$, if we have prior information on the minimum separation $\varepsilon^\star := \min_{i \neq j} \|c_i - c_j\|_2$ of $h^\star = (c_1, \dots, c_k)$ and on $R^\star := \max_l \|c_l\|_2$ then it is enough to choose the scale parameter $s \leq \varepsilon^\star(4\sqrt{\log(ek)})^{-1}$ and the sketch size large enough (with logarithmic dependency on $R^\star/\varepsilon^\star$) to ensure that $d(\pi^\star, \mathcal{H}) = 0$. Note that this holds with the sample space $\mathcal{Z} = \mathbb{R}^d$, i.e., we only restrict the optimization of the proxy $R_{\texttt{clust.}}(\mathbf{y}, h)$, *not the data* to centers in the Euclidean ball of radius $R \geq R^\star$, $\mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R)$.

Second, as we now show with a focus on $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$, even when $h^\star$ is not separated the term $d(\pi^\star, \mathcal{H})$ can remain under control. This comes from a combination of two main facts: 1) the risk for k-means and k-medians varies gently with respect to a natural distance between hypotheses; and 2) the distance between an arbitrary $h = (c_1, \dots, c_k)$ and the closest separated one is controlled. To formalize these facts we introduce the following notation:

Definition 3.6. *Given $\mathbf{c} = (c_1, \dots, c_k)$ and $\mathbf{c}' = (c'_1, \dots, c'_k)$ two k-tuples with $c_i, c'_j \in \mathbb{R}^d$, denote*

$$d(c_i \| \mathbf{c}') := \min_{1 \leq j \leq k} \|c_i - c'_j\|_2; \qquad d(\mathbf{c} \| \mathbf{c}') := \max_{1 \leq i \leq k} d(c_i \| \mathbf{c}'). \tag{30}$$

Since $d(\cdot\|\cdot)$ is not symmetric we define $d(\mathbf{c}, \mathbf{c}') := \max(d(\mathbf{c}\|\mathbf{c}'), d(\mathbf{c}'\|\mathbf{c}))$.

Definition 3.7. *For $\mathbf{c} = (c_1, \dots, c_k)$ with $c_i \in \mathbb{R}^d$, the (index) set of "ε-isolated" centroids of $\mathbf{c}$ (ignoring repetitions) is denoted $I_\varepsilon(\mathbf{c}) := \big\{i : 1 \leq i \leq k; \forall j \neq i : c_j = c_i \text{ or } \|c_i - c_j\|_2 \geq \varepsilon\big\}$.*

The following lemmas are proved in Appendix C.7.

Lemma 3.8. *Consider $k \geq 2$, $h, h' \in \mathcal{H} = (\mathbb{R}^d)^k$, and π with integrable k-means (resp. k-medians) loss. With $p = 2$ for k-means (resp. $p = 1$ for k-medians) we have $\left|\mathcal{R}(\pi, h)^{1/p} - \mathcal{R}(\pi, h')^{1/p}\right| \leq d(h, h')$.*

Since $\pi^\star = P_{h^\star}\pi = \sum_{i=1}^k \alpha_i \delta_{c_i}$ satisfies $\mathcal{R}(\pi^\star, h^\star) = 0$, by Lemma 3.8 we get $\mathcal{R}^{1/p}(\pi^\star, h) \leq d(h^\star, h)$ for each $h \in \mathcal{H}$, with $p = 1$ for k-medians and $p = 2$ for k-means. Hence, with the constants C and C' from Lemma 3.2, we have

$$d_{\texttt{k-medians}}(\pi^\star, \mathcal{H}) \leq C \inf_{h \in \mathcal{H}} d(h^\star, h);$$
$$d_{\texttt{k-means}}(\pi^\star, \mathcal{H}) \leq \inf_{h \in \mathcal{H}} \left\{ d^2(h^\star, h) + C' d(h^\star, h) \right\}.$$

Lemma 3.9. *Given $\varepsilon \geq 0$ and $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}_{k,0,R}$, there exists $\mathbf{c}' \in \mathcal{H}_{k,\varepsilon,R}$ such that $d(\mathbf{c}, \mathbf{c}') < \varepsilon$, and such that all ε-isolated centroids of $\mathbf{c}$, $\{c_i, i \in I_\varepsilon(\mathbf{c})\}$, are centroids of $\mathbf{c}'$ (ε-isolated centroids of $\mathbf{c}$ appearing multiple times may appear only once in $\mathbf{c}'$.)*

Denoting $R^\star := \max_i \|c_i\|_2$, by Lemma 3.9, there exists $h_\varepsilon \in \mathcal{H}_{k,2\varepsilon,R^\star}$ such that $d(h^\star, h_\varepsilon) \leq 2\varepsilon$. A consequence is that if $\mathcal{H}_{k,2\varepsilon,R^\star} \subseteq \mathcal{H}$, then

$$d_{\texttt{k-medians}}(\pi^\star, \mathcal{H}) \leq C d(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}) \leq 2C\varepsilon;$$
$$d_{\texttt{k-means}}(\pi^\star, \mathcal{H}) \leq d^2(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}) + C' d(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}) \leq 4\varepsilon^2 + 2C'\varepsilon.$$

Of course these are worst-case bounds. In particular, observe (using the definition of the constants C and C' in Lemma 3.2) that for large k we have $C\varepsilon \gg R$ and $C'\varepsilon \gg R^2$. Better bounds can be obtained by taking more finely into account the possible "near" separation of $h^\star$ as well as the weights α_i of the non ε-isolated centroids.

Lemma 3.10. *Let $\pi^\star = \sum_{i=1}^k \alpha_i \delta_{c_i}$ be an (arbitrary) k-mixture of Diracs and $h^\star := (c_1, \ldots, c_k)$ its centroids. Denote $R^\star := \max_i \|c_i\|_2$, and $\bar{W}(\pi^\star, \varepsilon) := \sum_{i \notin I_\varepsilon(h^\star)} \alpha_i \in [0, 1]$ the weight of non-ε-isolated centroids. If $\mathcal{H}_{k,2\varepsilon,R^\star} \subseteq \mathcal{H}$, then for the k-medians or k-means risk we have*

$$d_{\texttt{k-medians}}(\pi^\star, \mathcal{H}) \leq C \min\left\{ \bar{W}(\pi^\star, 4\varepsilon) \cdot d(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}),\ \bar{W}(\pi^\star, 2\varepsilon) \cdot 2\varepsilon \right\}; \tag{31}$$
$$d_{\texttt{k-means}}(\pi^\star, \mathcal{H}) \leq \min\Big\{ \bar{W}(\pi^\star, 4\varepsilon) \cdot \left(d^2(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}) + C' d^2(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star}) \right), \bar{W}(\pi^\star, 2\varepsilon) \cdot \left(4\varepsilon^2 + 2C'\varepsilon \right) \Big\}. \tag{32}$$

The form of (31)-(32) illustrates that, when restricting possible clustering hypotheses h to be separated, the best restricted risk $\mathcal{R}(\pi^\star, h)$ can be smaller if the unrestricted optimal clustering $\pi^\star$ has centroids globally well-approximated by a set of separated centroids, or if $\pi^\star$ puts large weight on isolated centroids, with both effects possibly compounding.

3.3 Learning algorithm ?

For compressive clustering, learning in the sketched domain means addressing the minimization of the proxy $R_{\texttt{clust.}}(\mathbf{y}, h)$ over $h \in \mathcal{H}$, which is analogous to the classical finite-dimensional least squares problem under a sparsity constraint. The latter is NP-hard, yet, under RIP conditions, provably good and computationally efficient algorithms (either greedy or based on convex relaxations) have been derived [Foucart and Rauhut, 2012]. Remark that the classic (non-compressed) k-means problem by minimization of the empirical risk is also known to be NP-hard [Garey et al., 1982, Aloise et al., 2009] and that guarantees for approaches such as K-means++ [Arthur and Vassilvitskii, 2007] are only in expectation and with a logarithmic sub-optimality factor.

It was shown practically in [Keriven et al., 2017] that a heuristic based on orthogonal matching pursuit (which neglects the separation and boundedness constraint associated to the class $\mathcal{H}_{k,2\varepsilon,R}$) is empirically able to recover sums of Diracs from sketches of the appropriate size. It must be noted that recovering sums of Diracs from Fourier observations has been studied in the case of regular low frequency measurements. In this problem, called super-resolution, it was shown that a convex proxy (convexity in the space of distributions using total variation regularization) for the non-convex optimization of the proxy $R_{\texttt{clust.}}(\mathbf{y}, h)$ is able to recover sufficiently separated Diracs [Candès and Fernandez-Granda, 2013, De Castro et al., 2016, Duval and Peyré, 2015]. In dimension one, an algorithmic approach to address the resulting convex optimization problem relies on semi-definite relaxation of dual optimization followed by root finding. Extension to dimension d and weighted random Fourier measurements in not straigthforward. Frank-Wolfe algorithms [Bredies and Pikkarainen, 2013] are more flexible for higher dimensions, and a promising direction for future research around practical sketched learning.

3.4 Improved sketch size guarantees?

Although Theorem 3.1 only provides guarantees when m is of the order of $k^2 d$ (up to logarithmic factors), the observed empirical phase transition pattern [Keriven et al., 2017] hints that m of the order of kd is in fact sufficient. This is intuitively what one would expect the "dimensionality" of the problem to be, since this is the number of parameters of the model $\mathfrak{S}^{\texttt{CT}}(\mathcal{H})$. In fact, as the parameters live in the cartesian product of k balls of radius R in $\mathbb{R}^d$ and the "resolution" associated to the separation assumption is ε, a naive approach to address the problem would consist in discretizing the parameter space into $N = \mathcal{O}\big((R/\varepsilon)^d\big)$ bins. Standard intuition from compressive sensing would suggest a sufficient number of measures m of the order of $k \log N = \mathcal{O}\big(kd \log \frac{R}{\varepsilon}\big)$. We leave a possible refinement of our analysis, trying to capture the empirically observed phase transition, for future work.

4 Guarantees for Compressive Gaussian Mixture Modeling

We consider Gaussian Mixture Modeling on the sample space $\mathcal{Z} = \mathbb{R}^d$, with k Gaussian components with *fixed, known invertible covariance* matrix $\mathbf{\Sigma} \in \mathbb{R}^d$. Denoting $\pi_c = \mathcal{N}(c, \mathbf{\Sigma})$, an hypothesis $h = (c_1, ..., c_k, \alpha_1, ..., \alpha_k)$ contains the means and weights of the components of a Gaussian Mixture Model (GMM) denoted $\pi_h = \sum_{l=1}^k \alpha_l \pi_{c_l}$, with $c_l \in \mathbb{R}^d$ and $\alpha \in \mathbb{S}_{k-1}$. The loss function for a density fitting problem is the negative log-likelihood: $\ell(x, h) = -\log \pi_h(x)$, and correspondingly the risk is $\mathcal{R}_{\texttt{GMM}}(\pi, h) = \mathbb{E}_{X\sim\pi}(-\log \pi_h(X))$. When π has a density with respect to the Lebesgue measure, and if this density admits a well-defined differential entropy,

$$\mathrm{H}(\pi) := \mathbb{E}_{X\sim\pi} - \log \pi(X), \tag{33}$$

the risk can be written $\mathcal{R}_{\texttt{GMM}}(\pi, h) = \mathrm{KL}(\pi||\pi_h) + \mathrm{H}(\pi)$ with $\mathrm{KL}(\pi||\pi') := \mathbb{E}_{X\sim\pi} \log \frac{\pi(X)}{\pi'(X)}$ the Kullback-Leibler divergence, see, e.g., [Cover and Thomas, 1991, Chapter 9]. For any distribution π with integrable GMM loss class, there exists an unconstrained global GMM risk minimizer $h^\star \in \mathbb{R}^{dk} \times \mathbb{S}_{k-1}$ of $\mathcal{R}_{\texttt{GMM}}(\pi, h)$ (see Section D.1 for a proof).

Model set $\mathfrak{S}^{\texttt{ML}}(\mathcal{H})$ and best hypothesis for $\pi \in \mathfrak{S}^{\texttt{ML}}(\mathcal{H})$. A natural model set for density fitting maximum log likelihood is precisely the model of all parametric densities:

$$\mathfrak{S}^{\texttt{ML}}(\mathcal{H}) := \{\pi_h : h \in \mathcal{H}\}. \tag{34}$$

A fundamental property of the Kullback-Leibler divergence is that $\mathrm{KL}(\pi||\pi') \geq 0$ with equality if, and only if $\pi = \pi'$. Hence, for any distribution $\widetilde{\pi} = \pi_{h_0}$ in the model set $\mathfrak{S}^{\texttt{ML}}(\mathcal{H})$, the optimum of minimization (14) (with $\varepsilon' = 0$) is $\hat{h} = h_0$ as it corresponds up to an offset independent of h to minimizing $\mathrm{KL}(\widetilde{\pi}||\pi_h)$.

Separation assumption. Similar to the compressive clustering framework case of Section 3, we enforce a minimum separation between the means of the components of a GMM. We denote

$$\mathcal{H}_{k,\varepsilon,R} = \left\{ (c_1, ..., c_k, \alpha_1, ..., \alpha_k) : c_l \in \mathbb{R}^d,\ \|c_l\|_{\mathbf{\Sigma}} \leq R,\ \min_{c_l \neq c_{l'}} \|c_l - c_{l'}\|_{\mathbf{\Sigma}} \geq \varepsilon,\ (\alpha_1, \ldots, \alpha_k) \in \mathbb{S}_{k-1} \right\}, \tag{35}$$

where

$$\|c\|_{\mathbf{\Sigma}} := \sqrt{c^T \mathbf{\Sigma}^{-1} c} \tag{36}$$

is the Mahalanobis norm associated to the known covariance $\mathbf{\Sigma}$.

Choice of feature function: random Fourier features. Compressive learning of GMMs with random Fourier features has been recently studied [Bourrier et al., 2013, Keriven et al., 2016]. Unlike compressive clustering we do not need to define a reweighted version of the Fourier features, and we directly sample m frequencies $\omega_1, \ldots, \omega_m$ in $\mathbb{R}^d$ i.i.d from the distribution with density

$$\Lambda = \Lambda_s = \mathcal{N}(0, s^{-2}\mathbf{\Sigma}^{-1}), \tag{37}$$

with scale parameter $s > 0$. Define the associated feature function $\Phi : \mathbb{R}^d \to \mathbb{C}^m$:

$$\Phi(x) := \frac{1}{\sqrt{m}} \left[e^{\jmath \omega_j^T x} \right]_{j=1,\ldots,m}. \tag{38}$$

Learning from a sketch by minimizing a proxy for the risk. Given sample points $x_1, \ldots, x_n$ in $\mathbb{R}^d$, a sketch $\mathbf{y}$ can be computed as in (2), i.e., as a sampling of the conjugate of the empirical characteristic function [Feuerverger and Mureika, 1977] of the distribution π of X. The characteristic function of a Gaussian $\pi_c = \mathcal{N}(c, \mathbf{\Sigma})$ has a closed form expression hence, with the operator $\mathcal{A}$ defined in (6), we have

$$\Psi(c) := \mathbb{E}_{X \sim \mathcal{N}(c,\mathbf{\Sigma})} \Phi(X) = \mathcal{A}(\pi_c) = \frac{1}{\sqrt{m}} \left[e^{\jmath \omega_j^T c} e^{-\frac{1}{2}\omega_j^T \mathbf{\Sigma} \omega_j} \right]_{j=1,\ldots,m}.$$

Then, given a sketch vector $\mathbf{y}$ and a hypothesis class $\mathcal{H}$, finding near minimizers of (13) and (14) in Theorem 2.2 corresponds to finding a (near) minimizer $\hat{h} = (\hat{c}_1, \ldots, \hat{c}_k, \hat{\alpha}_1, \ldots, \hat{\alpha}_k) \in \mathcal{H}$ of the proxy

$$R_{\texttt{GMM}}(\mathbf{y}, h) := \left\| \mathbf{y} - \sum_{i=1}^{k} \alpha_i \Psi(c_i) \right\|_2. \tag{39}$$

The following guarantees are proved in Appendix D.4 jointly with Theorem 3.1.

Theorem 4.1. *Consider Φ as in (38) where the ω_j are drawn according to (37) with scale $s \geq 1$. Given an integer $k \geq 1$ define $\varepsilon := 4\sqrt{(2+s^2)\log(ek)}$ and consider $R \geq \varepsilon$.*

1. *There is a universal constant $C > 0$ such that, for any $\zeta, \delta \in (0,1)$, when the sketch size satisfies*

$$m \geq C\delta^{-2} \cdot k \cdot \left[kd \cdot \left(\frac{d}{s^2} + 1 + \log(kRs) + \log(1/\delta) \right) + \log(1/\zeta) \right] \cdot \min(\log^2(ek), s^2 \log(ek)) \cdot (1 + 2/s^2)^{d/2}. \tag{40}$$

with probability at least $1 - \zeta$ on the draw of $(\omega_j)_{j=1}^m$ the operator $\mathcal{A}$ induced by Φ satisfies

$$1 - \delta \leq \frac{\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} \leq 1 + \delta, \qquad \forall \tau, \tau' \in \mathfrak{S}^{\texttt{ML}}(\mathcal{H}_{k,2\varepsilon,R}). \tag{41}$$

2. *Consider any samples $x_i \in \mathbb{R}^d$, $1 \leq i \leq n$ (represented by the empirical distribution $\hat{\pi}_n$) and any probability distribution π on $\mathbb{R}^d$ that is both $\mathcal{L}(\mathcal{H})$-integrable and $\{\Phi\}$-integrable.*
Consider a constrained class $\mathcal{H} \subseteq \mathcal{H}_{k,2\varepsilon,R} \subseteq \overline{\mathcal{H}} := (\mathbb{R}^d)^k \times \mathbb{S}_{k-1}$. Denote $h^\star \in \arg\min_{h \in \overline{\mathcal{H}}} \mathcal{R}_{\texttt{GMM}}(\pi, h)$, $\pi^\star := \pi_{h^\star}$, and consider $\hat{h} \in \mathcal{H}$ and $\nu, \nu' > 0$ such that

$$R_{\texttt{GMM}}(\mathbf{y}, \hat{h}) \leq (1+\nu) \inf_{h \in \mathcal{H}} R_{\texttt{GMM}}(\mathbf{y}, h) + \nu'. \tag{42}$$

with the proxy $R_{\texttt{GMM}}(\mathbf{y}, \cdot)$ defined in (39) and the sketch vector $\mathbf{y} := \frac{1}{n}\sum_{i=1}^n \Phi(x_i) = \mathcal{A}(\hat{\pi}_n)$.

If (41) holds then the excess risk of $\hat{h}$ with respect to $h^\star$ satisfies

$$\begin{aligned}\mathrm{KL}(\pi||\pi_{\hat{h}}) - \mathrm{KL}(\pi||\pi_{h^\star}) = \Delta\mathcal{R}_{h^\star}(\pi, \hat{h}) \leq & (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 \\ & + D_{h^\star}^{\overline{\mathcal{H}}}(\pi\|\pi^\star) + d(\pi^\star, \mathcal{H}) + C_{\mathcal{A}}\nu' \end{aligned} \tag{43}$$

where $C_{\mathcal{A}} \leq 46\sqrt{k/(1-\delta)}R^2\left(1 + 2/s^2\right)^{d/4}$ and

$$d(\pi^\star, \mathcal{H}) := \inf_{\tau \in \mathfrak{S}^{\texttt{ML}}(\mathcal{H})} \left\{ \sup_{h \in \mathcal{H}} (\mathrm{KL}(\pi^\star||\pi_h) - \mathrm{KL}(\tau||\pi_h)) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 \right\}. \tag{44}$$

Remark 4.2. *Note that this holds with the sample space $\mathcal{Z} = \mathbb{R}^d$, i.e., we only restrict the means of the GMM,* not the data*, to the ball of radius R, $\mathcal{B}_{\mathbb{R}^d, \|\cdot\|_{\mathbf{\Sigma}}}(0, R)$.*

The first term in the bound (43) is a statistical error term that is easy to control since (38) implies $\|\Phi(x)\|_2 = 1$ for each x. By the vectorial Hoeffding's inequality [Pinelis, 1992], for i.i.d. samples x_i drawn according to π, with high probability w.r.t. data sampling it holds that $C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2$ is of the order of at most $\left(1 + 2/s^2\right)^{d/4}\sqrt{k}R^2/\sqrt{n}$. To reach a given precision $\xi > 0$ we thus need $n \gtrsim \xi^{-2}\left(1 + 2/s^2\right)^{d/2} kR^4$ training samples. Notice that when s^2 is of the order of d this is of the order of $\xi^{-2}kR^4$. However $\left(1 + 2/s^2\right)^{d/2} \leq e^{d/s^2}$ can grow exponentially with d when s^2 is of order one, potentially requiring n to grow exponentially with d to have a small statistical error.

The second term $\|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2$ and the third one $D_{h^\star}^{\overline{\mathcal{H}}}(\pi\|\pi^\star)$ measure a modeling error, as they vanish when π belongs to the considered family of Gaussian mixtures. The second term can be controlled using Pinsker's inequality $\|\pi - \pi'\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(\pi||\pi')}$ [Fedotov et al., 2003]. Considering $\Phi_{\mathbf{u}}(x) := \langle \Phi(x), \mathbf{u} \rangle$ where $\mathbf{u} \in \mathbb{R}^m$ satisfies $\|\mathbf{u}\|_2 \leq 1$, we have $|\Phi_{\mathbf{u}}(x)| \leq \|\Phi(x)\|_2 = 1$ for all x. By definition of the total variation norm it follows that

$$\begin{aligned}\|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 = \sup_{\mathbf{u} \in \mathbb{R}^m, \|\mathbf{u}\|_2 \leq 1} \langle \mathcal{A}(\pi) - \mathcal{A}(\pi^\star), \mathbf{u} \rangle &= \sup_{\mathbf{u} \in \mathbb{R}^m, \|\mathbf{u}\|_2 \leq 1} \mathbb{E}_{X \sim \pi}\Phi_{\mathbf{u}}(x) - \mathbb{E}_{X \sim \pi^\star}\Phi_{\mathbf{u}}(x) \\ &\leq \|\pi - \pi^\star\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(\pi||\pi^\star)}.\end{aligned}$$

As $\mathcal{R}_{\texttt{GMM}}(\pi, h^\star)$ is, up to an additive offset, equal to $\mathrm{KL}(\pi||\pi^\star)$, this is reminiscent of the type of distribution free control obtained for clustering using [Gribonval et al., 2021, Lemma 3.4] Whether $D_{h^\star}^{\overline{\mathcal{H}}}(\pi\|\pi^\star)$ vanishes as in compressive clustering (cf [Gribonval et al., 2021, Lemma 3.2] and Lemma D.6 in the appendix) is an interesting question left to further work.

As in compressive clustering the fourth term, $d(\pi^\star, \mathcal{H})$, is a measure of distance of the best (unconstrained) gaussian mixture model to the considered constrained hypothesis class. Controlling this term as was done for compressive clustering in Lemma 3.2 would require further investigations.

Separation assumption. Given the scale parameter $s \geq 1$ and the number of Gaussians k, Theorem 4.1 sets a separation condition ε sufficient to ensure compressive statistical learning guarantees with the proposed sketching procedure, as well as a sketch size driven by M_s. Contrary to the case of Compressive Clustering, one cannot target an arbitrary small separation as for any value of s we have $\varepsilon \geq 4\sqrt{2\log(ek)}$. Reaching guarantees for a level of separation $\mathcal{O}\Big(\sqrt{\log(ek)}\Big)$ requires choosing s of the order of one. As we have just seen, this may require exponentially many training samples to reach a small estimation error, which is not necessarily surprising as such a level of separation is smaller than generally found in the literature [see e.g. Achlioptas and McSherry, 2005, Dasgupta and Schulman, 2000, Vempala and Wang, 2004]. For larger values of the scale parameter s, the separation required for our results to hold is larger.

Sketch size. Contrary to the case of Compressive Clustering (cf Theorem 3.1), the choice of the scale parameter s also impacts the sketch size required for the guarantees of Theorem 4.1 to hold. Choosing $s^2 = 2$ we get ε^2 of the order of $\log(ek)$, and (40) holds as soon as (with a universal numerical constant C that may vary from line to line below)

$$m \geq C\delta^{-2} \cdot 2^{d/2} \cdot k \cdot \{kd \cdot [d + \log k + \log R + \log(1/\delta)] + \log(1/\zeta)\} \cdot \log(ek).$$

Choosing $s^2 = d$ we get ε^2 of the order of $d\log(ek)$, and (40) holds as soon as

$$m \geq C\delta^{-2} \cdot k \cdot \{kd \cdot [1 + \log(kd) + \log(R) + \log(1/\delta)] + \log(1/\zeta)\} \cdot \log(ek) \min(\log(ek), d).$$

Choosing $s^2 = d/\log(ek)$ we get ε^2 of the order of $d + \log k$, and (40) holds as soon as

$$m \geq C\delta^{-2} \cdot k^2 \cdot \{kd \cdot [1 + \log(kd) + \log(R) + \log(1/\delta)] + \log(1/\zeta)\} \cdot \min(\log^2(ek), d).$$

Choosing $s^2 \gg d$ does not seem to pay off.

Tradeoffs. Overall we observe a trade-off between the required sketch size, the required separation of the means in the considered class of GMMs, and the sample complexity. When the scale parameter s decreases, higher frequencies are sampled (or, equivalently, the spatial kernel is more localized), and the required separation of means decreases. As a price, a larger number of sampled frequencies is required, and the sketch size increases as well as the factor $C_{\mathcal{A}}$. We give some particular values for

Scale s^2	Separation ε	Estimation error factor $C_{\mathcal{A}}$	Sketch size m
d	$\sqrt{d\log(ek)}$	$\sqrt{k}R^2$	$k^2 d \cdot \log(ekdR)\log^2(ek)$
$\frac{d}{\log(ek)}$	$\sqrt{d + \log(ek)}$	$k\sqrt{k}R^2$	$k^3 d \cdot \log(ekdR)\log^2(ek)$
2	$\sqrt{\log(ek)}$	$2^{d/2}\sqrt{k}R^2$	$k^2 d^2 \cdot 2^{d/2} \cdot (1 + \log(kR)/d)\log(ek)$

Table 1: Some tradeoffs between separation assumption, estimation error factor, and sketch size guarantees obtained using Theorem 4.1 for various values of the scale parameter s^2 of the frequency distribution (37). Each expression gives an order of magnitude up to universal numerical factors and factors depending only on δ and ζ.

s in Table 1. The regime $s^2 = 2$ may be useful to resolve close Gaussians in moderate dimensions (typically $d \leq 10$) where the factor $2^{d/2}$ in sample complexity and sketch size remains tractable.

Learning algorithm and improved sketch size guarantees? Again, although Theorem 4.1 only provides guarantees when m exceeds the order of $k^2 d$ (up to logarithmic factors, and for the most favorable choice of scale parameter s with the strongest separation constraints), the observed empirical phase transition pattern [Keriven et al., 2018] (using an algorithm to adress the optimization of (39) with a greedy heuristic) suggests that m of the order of kd, i.e. of the order of the number of unknown parameters, is in fact sufficient. Also, while Theorem 4.1 only handles mixtures of Gaussians with fixed known covariance matrix, the same algorithm has been observed to behave well for mixtures of Gaussians with unknown diagonal covariance.

5 Establishing the RIP for general mixture models

To establish the main results of the previous sections, Theorem 3.1 and Theorem 4.1, we will prove that the main assumption (11) of Theorem 2.2 holds with high probability. As recalled in Section 5.1 below (see Theorem 5.1), this can be achieved using the general approach described in [Gribonval et al., 2021, Section 5] relating random features and *kernel mean embeddings* of probability distributions, and using the notion of a *normalized secant set*. As the models sets appearing in Theorem 3.1 and Theorem 4.1 are mixture models (mixtures of k Dirac, or mixtures of k Gaussians), we develop in Section 5.2 tools for generic mixture models, introducing the notion of *(separated) dipole* and that of *mutual coherence* of separated dipoles.

5.1 Ingredients to establish the LRIP for randomized sketching

Considering a parameterized family of (real- or complex-valued) measurable functions $\mathcal{F} := \{\phi_\omega\}_{\omega\in\Omega}$ over $\mathcal{Z}$ and a probability distribution Λ over the parameter set Ω (often $\Omega \subseteq \mathbb{R}^d$), the random feature functions we consider are defined by drawing ω_j, $1 \leq j \leq m$, *i.i.d.* from the distribution Λ and defining

$$\Phi(x) := \tfrac{1}{\sqrt{m}}\big(\phi_{\omega_j}(x)\big)_{j=1,m}. \tag{45}$$

The expectation of $\langle \Phi(x), \Phi(x')\rangle = \frac{1}{m}\sum_{j=1}^m \phi_{\omega_j}(x)\overline{\phi_{\omega_j}(x')}$ defines a kernel

$$\kappa(x,x') := \mathbb{E}_{\omega\sim\Lambda}\phi_\omega(x)\overline{\phi_\omega(x')} \tag{46}$$

as well as the corresponding *mean embedding* kernel [Sriperumbudur et al., 2010] for probability distributions,

$$\kappa(\pi,\pi') := \mathbb{E}_{X\sim\pi}\mathbb{E}_{X'\sim\pi'}\kappa(X,X'), \tag{47}$$

and the associated Maximum Mean Discrepancy (MMD) metric

$$\|\pi-\pi'\|_\kappa := \sqrt{\kappa(\pi,\pi) - 2\mathrm{Re}(\kappa(\pi,\pi')) + \kappa(\pi',\pi')}. \tag{48}$$

By construction $\|\pi-\pi'\|_\kappa^2$ is the expectation (with respect to the draw of ω_j, $1\leq j\leq m$) of

$$\|\mathcal{A}(\pi)-\mathcal{A}(\pi')\|_2^2 = \frac{1}{m}\sum_{j=1}^m \big|\mathbb{E}_{X\sim\pi}\phi_{\omega_j}(X) - \mathbb{E}_{X'\sim\pi'}\phi_{\omega_j}(X')\big|^2.$$

A quantity of interest, given a model set $\mathfrak{S}$, is a *concentration function* $t\mapsto c_\kappa(t)\in(0,\infty]$ such that

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\tau)-\mathcal{A}(\tau')\|_2^2}{\|\tau-\tau'\|_\kappa^2} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{m}{c_\kappa(t)}\right), \qquad \forall \tau,\tau'\in\mathfrak{S}, \quad \forall t>0, \qquad \forall m\geq 1. \tag{49}$$

The *normalized secant set* of the model set $\mathfrak{S}$ with respect to a kernel κ is the following subset of the set of finite signed measures (see [Gribonval et al., 2021, Appendix A.2]):

$$\mathcal{S}_\kappa = \mathcal{S}_\kappa(\mathfrak{S}) := \left\{ \frac{\tau - \tau'}{\|\tau - \tau'\|_\kappa} : \tau, \tau' \in \mathfrak{S}, \|\tau - \tau'\|_\kappa > 0 \right\}. \tag{50}$$

Given a function class $\mathcal{G}$ of measurable functions $g : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$, the *radius* of a subset $\mathcal{E}$ of finite signed measures is denoted

$$\|\mathcal{E}\|_{\mathcal{G}} := \sup_{\mu \in \mathcal{E}} \|\mu\|_{\mathcal{G}} = \sup_{\mu \in \mathcal{E}} \sup_{g \in \mathcal{G}} \left| \int g d\mu \right|. \tag{51}$$

Of particular interest will be $\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}$ and $\|\mathcal{S}_\kappa\|_{\mathcal{F}}$. The *covering number* $\mathrm{N}(d(\cdot,\cdot), S, \delta)$ of a set S with respect to a (pseudo)metric[4] $d(\cdot,\cdot)$ is the minimum number of closed balls of radius δ with respect to $d(\cdot,\cdot)$ with centers in S needed to cover S. We can now recall [Gribonval et al., 2021, Theorem 5.7]:

Theorem 5.1. *Consider $\mathcal{F} := \{\phi_\omega\}_{\omega \in \Omega}$ a family of real or complex-valued functions on $\mathcal{Z}$, Λ a probability distribution on Ω, Φ the associated random feature function and κ the corresponding kernel. Consider the pseudometric on $\mathcal{F}$-integrable probability distributions*[5]

$$d_{\mathcal{F}}(\pi, \pi') := \sup_{\omega \in \Omega} \left| |\mathbb{E}_{X \sim \pi} \phi_\omega(X)|^2 - |\mathbb{E}_{X' \sim \pi'} \phi_\omega(X')|^2 \right|. \tag{52}$$

Consider a model set $\mathfrak{S}$ and $\mathcal{S}_\kappa$ its normalized secant set. Assume the pointwise concentration function $c_\kappa(\delta)$ satisfying (49) exists. For $0 < \delta, \zeta < 1$, if

$$m \geq c_\kappa(\delta/2) \cdot \log\Big(2\mathrm{N}(d_{\mathcal{F}}, \mathcal{S}_\kappa, \delta/2)/\zeta\Big), \tag{53}$$

then, with probability at least $1 - \zeta$ on the draw of $(\omega_j)_{j=1}^m$, the operator $\mathcal{A}$ induced by Φ (cf (6)) satisfies

$$1 - \delta \leq \frac{\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} \leq 1 + \delta, \qquad \forall \tau, \tau' \in \mathfrak{S}. \tag{54}$$

When (54) holds, the LRIP (11) holds with constant $C_{\mathcal{A}} := \frac{\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$ and $\eta = 0$.

5.2 Separated mixtures models, dipoles, and mutual coherence

In Theorems 3.1 and 4.1, the random feature map Φ is made of (weighted) random Fourier features, leading to a shift-invariant kernel κ, and the considered model set is a mixture of Diracs (resp. of Gaussians) satisfying a certain separation condition. To prove these theorems using Theorem 5.1, our main goal is to bound the radius of the normalized secant set, $\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}$, as well as the concentration function $c_\kappa(t)$ (see (49)) and the covering numbers of $\mathcal{S}_\kappa$ (see (50)) with respect to the pseudometric (52). As the distance $\|\tau - \tau'\|_\kappa$ is the denominator of all these expressions, most difficulties arise when $\|\tau - \tau'\|_\kappa$ is small ($\tau, \tau' \in \mathfrak{S}$ get "close" to each other) and we primarily have to control the ratio $\|\tau - \tau'\|/\|\tau - \tau'\|_\kappa$ for various norms when $\|\tau - \tau'\|_\kappa \to 0$. In this section, we develop a framework to control these quantities when the model $\mathfrak{S}$ is a mixture model, which covers both mixtures of Diracs and mixtures of Gaussians.

We consider a given parametrized family of base distributions $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ where $\mathbf{\Theta}$ is a parameter set (typically a subset of a finite-dimensional vector space), ϱ is a metric on $\mathbf{\Theta}$, and $\varphi : \theta \in \mathbf{\Theta} \mapsto \varphi(\theta) = \pi_\theta$ is an *injective* map defining a family of probability distributions (e.g. a family of Diracs

[4] Further reminders on metrics, pseudometrics, and covering numbers are given in [Gribonval et al., 2021, Appendix A].

[5] In fact, we consider the extension of $d_{\mathcal{F}}$ to finite signed measures, see Appendix A.2 in [Gribonval et al., 2021].

or of Gaussians). In statistical terms, $\mathcal{T}$ is an identifiable statistical model whose parameter space is equipped with a metric, and π_θ is a (probability) measure on the sample space $\mathcal{Z}$. We define 2-separated k-mixtures from $\mathcal{T}$ as

$$\mathfrak{S}_k(\mathcal{T}) := \left\{ \tau = \sum_{l=1}^{\ell} \alpha_l \pi_{\theta_l} : \ell \leq k,\ \alpha_l > 0,\ \sum_{l=1}^{\ell} \alpha_l = 1,\ \theta_l \in \mathbf{\Theta},\ \varrho(\theta_l, \theta_{l'}) \geq 2\ \forall l \neq l' \leq \ell \right\}. \tag{55}$$

Remark 5.2. *In the case of Diracs $\pi_\theta = \delta_\theta$, with $\varrho(\theta, \theta') = \|\theta - \theta'\|_2/\varepsilon$, $\mathfrak{S}_k(\mathcal{T})$ is the set of mixtures of k pairwise 2ε-separated Diracs considered in Section 3. For Gaussians $\pi_\theta = \mathcal{N}(\theta, \mathbf{\Sigma})$, $\varrho(\theta, \theta') := \|\theta - \theta'\|_{\mathbf{\Sigma}}/\varepsilon$, we obtain the set of 2ε-separated Gaussian mixtures considered in Section 4.*

The notion of *dipoles* will turn out to be particularly useful in our analysis.

Definition 5.3 (Dipoles, separation)**.** *A finite signed measure*[6] *ν is a* **dipole** *with respect to $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ if it admits a decomposition as $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$ where $\theta_1, \theta_2 \in \mathbf{\Theta}$, $\varrho(\theta_1, \theta_2) \leq 1$ and $\alpha_i \geq 0$ for $i = 1, 2$. The coefficients α_i's are not necessarily normalized to 1, and any of them can be put to 0 to yield a* monopole *as a special case. Two dipoles ν, ν' are* **1-separated** *if they admit a decomposition $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$, $\nu' = \alpha'_1 \pi_{\theta'_1} - \alpha'_2 \pi_{\theta'_2}$ as above such that $\varrho(\theta_i, \theta'_j) \geq 1$ for all $i, j \in \{1, 2\}$.*

The relevance of the notion of separated dipoles to handle the secant of separated mixtures is captured in the following decomposition lemma:

Lemma 5.4. *If $\tau, \tau' \in \mathfrak{S}_k(\mathcal{T})$, then there exists $\ell \leq 2k$ nonzero dipoles $(\nu_l)_{1 \leq l \leq \ell}$ that are pairwise 1-separated and satisfy $\tau - \tau' = \sum_{l=1}^{\ell} \nu_l$.*

Proof. Using the 2-separation in τ and τ' and the triangle inequality, for the metric ϱ each parameter θ_i in τ is 1-close to *at most* one parameter θ'_j in τ', and 1-separated from all other components in both τ and τ'. Hence $\tau - \tau'$ can be decomposed into a sum of (at most) $2k$ dipoles (some of which may also be monopoles). □

As announced previously, we are interested in RIP inequalities with the kernel norm in the denominator. Correspondingly, it is natural to introduce the notion of *normalized* monopoles and dipoles, given a kernel κ and the associated mean map embedding. It will be convenient to make some basic assumptions on this kernel. For the following definitions, we only assume κ is a positive semi-definite (psd) kernel on $\mathcal{Z}$ with the associated kernel mean embedding defined by (47); the explicit representation in terms of random features is not needed.

Definition 5.5 (Locally characteristic kernel, normalized kernel)**.** *A psd kernel κ on $\mathcal{Z}$ (extended to probability distributions on $\mathcal{Z}$ via the kernel mean embedding* (47)*) is* locally characteristic *with respect to $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ if it satisfies the following two conditions:*

1. *$\|\pi_\theta\|_\kappa > 0$ for each $\theta \in \mathbf{\Theta}$;*
2. *$|\kappa(\pi_\theta, \pi_{\theta'})| < \|\pi_\theta\|_\kappa \|\pi_{\theta'}\|_\kappa$ for each $\theta \neq \theta' \in \mathbf{\Theta}$ such that $\varrho(\theta, \theta') \leq 1$.*

Note that if κ is locally characteristic, then $\|\nu\|_\kappa > 0$ for any nonzero dipole.

Definition 5.6 (Normalized monopoles, normalized dipoles)**.** *The set of* **normalized dipoles** *induced by the base family $\mathcal{T}$ with respect to a locally characteristic kernel κ is denoted by*

$$\mathcal{D} = \mathcal{D}_\kappa(\mathcal{T}) := \left\{ \frac{\nu}{\|\nu\|_\kappa} : \nu \text{ is a nonzero dipole} \right\}. \tag{56}$$

[6] See Appendix A.2 in [Gribonval et al., 2021]

It contains as a particular subset the set of **normalized monopoles**

$$\mathcal{M} = \mathcal{M}_\kappa(\mathcal{T}) := \left\{ \nu_\theta := \frac{\pi_\theta}{\|\pi_\theta\|_\kappa} : \theta \in \mathbf{\Theta} \right\}. \tag{57}$$

Equipped with these notions we can define the mutual coherence and ℓ-coherence of a kernel.

Definition 5.7. *A psd kernel κ on $\mathcal{Z}$ has mutual coherence M with respect to $\mathcal{T}$ if: (a) it is locally characteristic with respect to $\mathcal{T}$; and (b) for each pair of normalized dipoles $\mu, \mu' \in \mathcal{D}_\kappa(\mathcal{T})$ that are 1-separated from each other, we have*[7]

$$|\kappa(\mu, \mu')| \leq M. \tag{58}$$

Given an integer $\ell > 0$ and a number $\zeta \in [0, 1]$, we say that a kernel κ has its ℓ-coherence with respect to $\mathcal{T}$ bounded by ζ if, for any dipoles $(\nu_l)_{1 \leq l \leq \ell}$ that are pairwise 1-separated and such that $\sum_{l=1}^{\ell} \|\nu_l\|_\kappa^2 > 0$, it holds

$$1 - \zeta \leq \frac{\left\| \sum_{l=1}^{\ell} \nu_l \right\|_\kappa^2}{\sum_{l=1}^{\ell} \|\nu_l\|_\kappa^2} \leq 1 + \zeta. \tag{59}$$

A crucial step in the analysis to come is the reduction from differences of k-mixtures to individual dipoles. To this end, the representation of Lemma 5.4 combined with the quasi-Pythagorean identity (59) will play a central role. The following result is a direct consequence of Gershgorin's disc lemma [see e.g. Foucart and Rauhut, 2012, Theorem 5.3] and establishes the link between mutual coherence and ℓ-coherence.

Lemma 5.8. *Consider a kernel κ with mutual coherence M with respect to $\mathcal{T}$. Then κ has ℓ-coherence bounded by $M(\ell - 1)$.*

Remark 5.9. *The reader familiar with sparse recovery will find this lemma highly reminiscent of the classical link between the coherence of a dictionary and its restricted isometry property [see e.g. Foucart and Rauhut, 2012, Theorem 5.13]. To handle incoherence in a continuous "off the grid" setting (such as mixtures of separated Diracs in Section 3, which also appear in super-resolution imaging scenarios [Candès and Fernandez-Granda, 2013, De Castro et al., 2016, Duval and Peyré, 2015]), the apparently new trick is to consider incoherence* between dipoles *rather than between monopoles.*

Conditions such that κ has low mutual coherence with respect to $\mathcal{T}$ will be given in Theorem 5.16.

5.3 From separated k-mixtures to dipoles

We turn to the ingredients delineated in Section 5.1 in order to establish the RIP for (separated) k-mixture models. Using the notions introduced in Section 5.2, the following results allow to control the various key quantities in terms of related notions defined by replacing the normalized secant set of k-mixtures with the simpler set $\mathcal{D}$ of normalized dipoles.

In the sequel we will generically assume to have fixed a base distribution family $\mathcal{T}$, a kernel κ, the associated normalized dipole and monopole sets $\mathcal{D} = \mathcal{D}_\kappa(\mathcal{T}), \mathcal{M} = \mathcal{M}_\kappa(\mathcal{T})$, an integer $k \geq 1$, the separated k-mixture model $\mathfrak{S} = \mathfrak{S}_k(\mathcal{T})$ and its normalized secant $\mathcal{S}_\kappa = \mathcal{S}_\kappa(\mathfrak{S})$ as introduced in the previous section. Our first result relates the radius of the normalized secant set with respect to any function family $\mathcal{G}$ to the corresponding radius of the set of dipoles.

[7] We properly define in Appendix A.2 of [Gribonval et al., 2021]the extension of the Mean Map Embedding to finite signed measures, to make sense of the notation $\kappa(\nu, \nu')$.

Theorem 5.10. *Assume the kernel κ has its $2k$-coherence with respect to $\mathcal{T}$ bounded by $\zeta \leq 3/4$. Let $\mathcal{G}$ be a real or complex-valued measurable function class over $\mathcal{Z}$. We have*

$$\|\mathcal{S}_\kappa\|_{\mathcal{G}} \leq \sqrt{8k} \cdot \|\mathcal{D}\|_{\mathcal{G}}. \tag{60}$$

Proof. First, by definition of $\|\mathcal{D}\|_{\mathcal{G}}$, we have $\|\nu\|_{\mathcal{G}} \leq \|\mathcal{D}\|_{\mathcal{G}} \cdot \|\nu\|_\kappa$ for any dipole ν. Let $\tau,\ \tau' \in \mathfrak{S}_k(\mathcal{T})$. Using Lemma 5.4 we write $\tau - \tau' = \sum_{i=1}^{\ell} \nu_i$ where $\ell \leq 2k$ and the ν_i's are dipoles that are pairwise 1-separated. By the triangle inequality and the Cauchy-Schwarz inequality we have

$$\|\tau - \tau'\|_{\mathcal{G}} \leq \sum_{i=1}^{\ell} \|\nu_i\|_{\mathcal{G}} \leq \|\mathcal{D}\|_{\mathcal{G}} \cdot \sum_{i=1}^{\ell} \|\nu_i\|_\kappa \leq \|\mathcal{D}\|_{\mathcal{G}} \cdot \sqrt{\ell} \left(\sum_{i=1}^{\ell} \|\nu_i\|_\kappa^2 \right)^{\frac{1}{2}}$$

By our assumption on the bounded $2k$-coherence of κ and since $\ell \leq 2k$ and $\zeta \leq 3/4$, we have

$$\|\tau - \tau'\|_{\mathcal{G}} \leq \frac{\|\mathcal{D}\|_{\mathcal{G}}}{\sqrt{1-\zeta}} \sqrt{\ell} \left\| \sum_{i=1}^{\ell} \nu_i \right\|_\kappa \leq 2\sqrt{2k} \cdot \|\mathcal{D}\|_{\mathcal{G}} \cdot \|\tau - \tau'\|_\kappa. \qquad \square$$

We now consider the random sketching operator: consider a family of functions $\mathcal{F} := \{\phi_\omega\}_{\omega \in \Omega}$, m parameters $(\omega_j)_{j=1}^m$ drawn i.i.d. according to some distribution Λ on Ω, $\mathcal{A}$ the operator induced (see (6)) by the feature function $\Phi(x) := \frac{1}{\sqrt{m}} \big(\phi_{\omega_j}(x)\big)_{j=1}^m$, and finally κ the associated average kernel, given by (46). For short, we call $(\mathcal{F}, \Lambda)$ a random feature family, and $\mathcal{A}, \kappa$ the induced (random) sketching operator and kernel.

Concerning the pointwise concentration function for this random sketching operator, by [Gribonval et al., 2021, Lemma 5.5], we have $\|\mathcal{S}_\kappa\|_{\mathcal{F}} \geq 1$, and the concentration function satisfies

$$c_\kappa(t) \leq 2t^{-2}(1 + t/3) \cdot \|\mathcal{S}_\kappa\|_{\mathcal{F}}^2, \qquad \forall t > 0. \tag{61}$$

Observe that this is based on a supremum control over the class $\mathcal{F}$, using the radius $\|\mathcal{S}_\kappa\|_{\mathcal{F}}$, and as such is independent of the choice of the distribution Λ over its index set. In settings such as Compressive Clustering with $d \gtrsim \log k$, sharper bounds on the concentration function can be obtained for mixture models when the considered kernel has low mutual coherence. In this situation, thanks to the separation assumption, *it is sufficient to properly control the moments wrt. Λ of normalized dipoles*, for which sharper bounds may be available.

For notational brevity, we extend by linearity the operator $\mathcal{A}$ to finite signed measures (in particular for normalized dipoles), and for any finite signed measure μ, we denote by $\langle \mu, f \rangle = \int f d\mu$ for an integrable function f. Proofs of the remaining results in this section are in Appendix B.

Theorem 5.11. *Consider a random feature family $(\{\phi_\omega\}_{\omega \in \Omega}, \Lambda)$ and the induced random sketching operator $\mathcal{A}$ and kernel κ. Assume κ has its $2k$-coherence with respect to $\mathcal{T}$ bounded by $\zeta \leq 3/4$.*

Assume there are $\gamma > 0, \lambda \geq 1$ such that, for each normalized dipole $\mu \in \mathcal{D}$:

$$\mathbb{E}_{\omega \sim \Lambda} \Big[|\langle \mu, \phi_\omega \rangle|^{2q} \Big] \leq \frac{q!}{2} \lambda \gamma^{q-1}, \qquad \textit{for each integer } q \geq 2. \tag{62}$$

Set $V := 16ek\gamma \log^2(4ek\lambda)$. For any $\mu \in \mathcal{S}_\kappa(\mathfrak{S}_k(\mathcal{T}))$ we have

$$\mathbb{P}\Big(\Big| \|\mathcal{A}(\mu)\|^2 - 1 \Big| \geq t \Big) \leq 2 \exp\left(-\frac{mt^2}{2V(1 + t/3)} \right), \qquad \textit{for each } t > 0. \tag{63}$$

Specific estimates of γ such that the moment bounds (62) hold for normalized dipoles will be given in Section 6 (Lemma 6.5) and completed in Section D where we gather all ingredients to prove Theorems 3.1 and 4.1 for Compressive Clustering and Compressive GMM.

Finally, the covering numbers (for $d_{\mathcal{F}}$) of the normalized secant set are also controlled by those (for $\|\cdot\|_{\mathcal{F}}$) of normalized dipoles.

Theorem 5.12. *Consider a random feature family $(\mathcal{F}, \Lambda)$ and the induced random sketching operator $\mathcal{A}$ and average kernel κ. Assume that κ is locally characteristic with respect to $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$.*

- *We have $\|\mathcal{D}\|_{\mathcal{F}} \geq 1$, and for each $\theta, \theta' \in \mathbf{\Theta}$ such that $\varrho(\theta, \theta') \leq 1$ and $\alpha, \alpha' \geq 0$*

$$\|\alpha\pi_\theta - \alpha'\pi_{\theta'}\|_\kappa \leq \|\alpha\pi_\theta - \alpha'\pi_{\theta'}\|_{\mathcal{F}} \leq \|\mathcal{D}\|_{\mathcal{F}}\|\alpha\pi_\theta - \alpha'\pi_{\theta'}\|_\kappa. \tag{64}$$

- *Assume the kernel κ has its $2k$-coherence with respect to $\mathcal{T}$ bounded by $\zeta \leq 3/4$, and consider $d_{\mathcal{F}}$ the pseudo-metric defined in (52). Then we have for each $\delta > 0$:*

$$\mathrm{N}(d_{\mathcal{F}}, \mathcal{S}_\kappa, \delta) \leq \left[\mathrm{N}\Big(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \tfrac{\delta}{64k\|\mathcal{D}\|_{\mathcal{F}}}\Big) \cdot \max\Big(1, \tfrac{256k\|\mathcal{D}\|_{\mathcal{F}}^2}{\delta}\Big)\right]^{2k}. \tag{65}$$

Gathering all the ingredients above together with the general Theorem 5.1 and Lemma 5.8 we obtain the following result.

Theorem 5.13. *Consider $\mathcal{F} := \{\phi_\omega\}_{\omega\in\Omega}$, Λ a probability distribution on Ω, and κ the induced average kernel. Assume that κ has mutual coherence M with respect to $\mathcal{T}$ and consider $k \geq 1$ such that $M(2k-1) \leq 3/4$. Assume that there are $C \geq 1$, $r > 0$ such that $\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \delta) \leq 2(C/\delta)^r$ for each $0 < \delta < 1$ and that there are $\gamma > 0, \lambda \geq 1$ such that*

$$\sup_{\mu\in\mathcal{D}} \mathbb{E}_{\omega\sim\Lambda}\Big[|\langle\mu, \phi_\omega\rangle|^{2q}\Big] \leq \frac{q!}{2}\lambda\gamma^{q-1}, \tag{66}$$

for every integer $q \geq 2$. For $0 < \delta, \zeta < 1$, if $(\omega_j)_{j=1}^m$ are drawn i.i.d. according to Λ and

$$m \geq 80 \cdot \delta^{-2} \cdot \min\Big(2e\gamma\log^2(4ek\lambda), \|\mathcal{D}\|_{\mathcal{F}}^2\Big) \cdot k \cdot \Big\{2k(r+1)\Big[\log(kC\|\mathcal{D}\|_{\mathcal{F}}^2) + \log(1024/\delta)\Big] + \log(2/\zeta)\Big\}, \tag{67}$$

then, with probability at least $1-\zeta$ on the draw of $(\omega_j)_{j=1}^m$, we have

$$1 - \delta \leq \frac{\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} \leq 1 + \delta, \qquad \forall \pi, \pi' \in \mathfrak{S}_k(\mathcal{T}). \tag{68}$$

where $\mathcal{A}$ is the operator induced by $\Phi(x) := \frac{1}{\sqrt{m}}\big(\phi_{\omega_j}(x)\big)_{j=1}^m$.

When (68) holds, the LRIP (11) holds with $C_{\mathcal{A}} := \frac{8\sqrt{2k}\|\mathcal{D}\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$ and $\eta = 0$ for each loss class $\mathcal{L}$.

5.4 Strongly characteristic kernels and associated controls

Theorem 5.13 notably involves two important quantities: the coherence M of the kernel, and the radiuses $\|\mathcal{D}\|_{\mathcal{G}}$, where $\mathcal{G} \in \{\Delta\mathcal{L}, \mathcal{F}\}$ and $\mathcal{D}$ is the set of normalized dipoles. These quantities can be controlled under some assumptions on $\mathcal{T}$ and κ which are essentially captured by a normalized version of the kernel, which we introduce now.

Definition 5.14. *Let $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ be a family of base distributions, and κ be a psd kernel on $\mathcal{Z}$ such that $\|\pi_\theta\|_\kappa > 0$ for each $\theta \in \mathbf{\Theta}$. We define the $\mathcal{T}$-normalized kernel $\overline{\kappa}$ on the parameter space $\mathbf{\Theta}$ as*

$$\overline{\kappa}(\theta, \theta') := \frac{\kappa(\pi_\theta, \pi_{\theta'})}{\|\pi_\theta\|_\kappa\|\pi_{\theta'}\|_\kappa}, \quad \theta, \theta' \in \mathbf{\Theta}. \tag{69}$$

It holds that $\overline{\kappa}(\theta, \theta) = 1$ for each $\theta \in \mathbf{\Theta}$, $|\overline{\kappa}(\theta, \theta')| \leq 1$ for every θ, θ', and κ is locally characteristic iff $|\overline{\kappa}(\theta, \theta')| < 1$ when $0 < \varrho(\theta, \theta') \leq 1$.

Given $c \in (0, 2]$ we say that the kernel κ is c-strongly locally characteristic *if it is real-valued and*

$$1 - \overline{\kappa}(\theta, \theta') \geq \frac{c}{2}\varrho^2(\theta, \theta'), \qquad \forall \theta, \theta' \in \mathbf{\Theta} \text{ such that } \varrho(\theta, \theta') \leq 1. \tag{70}$$

We note that kernels that locally decrease quadratically also appear naturally in sparse spikes recovery [Poon et al., 2020], where infinite-dimensional convex relaxations are employed to estimate sums of Diracs; like we do here for k-means/medians however through the non-convex problem (27). Concrete examples of such kernels will be given in Section 6, where typically $\varrho(\cdot,\cdot)$ is a simple Euclidean distance and κ is a Gaussian kernel.

Our first result relates $\mathcal{G}$-radiuses of the set of normalized dipoles $\mathcal{D}$ to those of the set of the normalized monopoles $\mathcal{M}$.

Theorem 5.15. *Consider a kernel κ that is c-strongly locally characteristic with respect to $\mathcal{T}$. For any function class $\mathcal{G}$ we have*

$$\|\mathcal{M}\|_{\mathcal{G}} \leq \|\mathcal{D}\|_{\mathcal{G}} \leq \|\mathcal{M}\|_{\mathcal{G}} + L_{\mathcal{G}}/\sqrt{c}, \tag{71}$$

with $L_{\mathcal{G}}$ the Lipschitz constant of $\theta \mapsto \nu_\theta := \pi_\theta / \|\pi_\theta\|_\kappa$ with respect to the metrics ϱ and $\|\cdot\|_{\mathcal{G}}$.

The proof is in Appendix B.4. The second result gives a concrete criterion to establish quantitatively that κ is c-strongly characteristic and has bounded mutual coherence.

Theorem 5.16. *Consider $\mathcal{T} = (\mathbf{\Theta}, \varrho, \psi)$ a family of base distributions, and κ a psd kernel on $\mathcal{Z}$. Assume that $\|\pi_\theta\|_\kappa > 0$ for each $\theta \in \mathbf{\Theta}$ and that the normalized kernek $\overline{\kappa}$ is of the form*

$$\overline{\kappa}(\theta, \theta') = K(\varrho(\theta, \theta')), \quad \forall \theta, \theta' \in \mathbf{\Theta}, \tag{72}$$

for a function $K : \mathbb{R}_+ \to \mathbb{R}_+$ such that $K(0) = 1$ and $0 \leq K(u) \leq 1 - \frac{cu^2}{2}, \forall u \in [0,1]$, with $0 < c \leq 2$. Then:

1. *The kernel κ is c-strongly locally characteristic with respect to $\mathcal{T}$.*
2. *If K is bounded and differentiable with bounded and Lipschitz derivative on $[1, \infty)$, and if there exists a mapping $\psi : \mathbf{\Theta} \mapsto \mathcal{H}$, with $\mathcal{H}$ some Hilbert space, such that $\varrho(\theta, \theta') := \|\psi(\theta) - \psi(\theta')\|_{\mathcal{H}}$, then the kernel κ has mutual coherence with respect to $\mathcal{T}$ bounded by*

$$M \leq \tfrac{4C}{\min(c,1)}, \tag{73}$$

with

$$C = C(K) := \max(K_{\max}, (2K'_{\max} + K''_{\max})), \tag{74}$$

where $K_{\max} := \sup_{u \geq 1} |K(u)|$, $K'_{\max} := \sup_{u \geq 1} |K'(u)|$, $K''_{\max} := \sup_{\substack{u,v \geq 1 \\ u \neq v}} \frac{|K'(u) - K'(v)|}{|u - v|}$.

The proof is in Appendix B.5.

6 Random Fourier Sketching with location-based mixtures

Given the prominent role of Random Fourier Features for Compressive Clustering and Compressive Gaussian Mixture Modeling, we now focus on this specific setting. Mixtures of Diracs / Gaussians belong to what we call *location-based mixture models*. Combined with a shift-invariant kernel on samples they yield a shift-invariant mean embedding, and we show that the assumptions of Theorem 5.13 and Theorem 5.16 are satisfied. We note that in [Poon et al., 2020], non-translation-invariant embeddings are treated with the same techniques as translation-invariant ones through a Riemannian geometry framework, which is an interesting path for future extensions.

6.1 Location-based mixtures and shift-invariant kernels

Much like the introduced notion of family of parametrized base distributions $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ is, in statistical terminology, an identifiable statistical model, the following definition specializes it to the case where the distributions in that collection are obtained by translation of a single reference distribution, which is generally called a location family.

Definition 6.1 (Location family, location-based mixtures)**.** *Consider $\|\cdot\|$ a norm on $\mathbb{R}^d$, π_0 a probability distribution on $\mathcal{Z} = \mathbb{R}^d$. For $\theta \in \mathbb{R}^d$, denote π_θ the distribution of $\theta + X$ when $X \sim \pi_0$ and consider the mapping $\varphi : \theta \mapsto \pi_\theta$. In statistical terms, given $\mathbf{\Theta} \subset \mathbb{R}^d$, $\mathcal{T} = (\mathbf{\Theta}, \|\cdot\|, \varphi)$ is a* location family*. We call $\mathfrak{S}_k(\mathcal{T})$, where $k \geq 1$, a* location-based mixture model*.*

Proposition 6.2. *Consider π_0 a probability distribution and κ a shift-invariant kernel on $\mathbb{R}^d$ such that $\|\pi_0\|_\kappa > 0$. Let $\mathcal{T}$ be a location family based on π_0. For each $\theta \in \mathbf{\Theta}$ we have $\|\pi_\theta\|_\kappa = \|\pi_0\|_\kappa$. There exists $\mathtt{K} : \mathbb{R}^d \mapsto \mathbb{R}$ such that $|\mathtt{K}(\theta)| \leq \mathtt{K}(0) = 1$ for every $\theta \in \mathbb{R}^d$ and*

$$\overline{\kappa}(\theta, \theta') = \mathtt{K}(\theta - \theta'), \quad \forall \theta, \theta' \in \mathbf{\Theta}. \tag{75}$$

By a standard abuse of notation we also denote $\overline{\kappa}$ the function $\mathtt{K}$, so that $\overline{\kappa}(\theta, \theta') = \overline{\kappa}(\theta - \theta')$.

Proof. Since κ is shift-invariant, there is g such that $\kappa(x, x') = g(x - x')$ for each $x, x' \in \mathcal{Z}$, hence

$$\begin{aligned}\kappa(\pi_\theta, \pi_{\theta'}) &= \mathbb{E}_{X\sim\pi_\theta}\mathbb{E}_{X'\sim\pi_{\theta'}}\kappa(X, X') = \mathbb{E}_{X\sim\pi_0}\mathbb{E}_{X'\sim\pi_0}\kappa(X + \theta, X' + \theta') \\ &= \mathbb{E}_{X\sim\pi_0}\mathbb{E}_{X'\sim\pi_0} g(\theta - \theta' + X - X')\end{aligned}$$

only depends on $\theta - \theta'$. As a result, there is a function $G : \mathbb{R}^d \to \mathbb{R}$ such that $\kappa(\pi_\theta, \pi_{\theta'}) = G(\theta - \theta')$. In particular, $\|\pi_\theta\|_\kappa^2 = \kappa(\pi_\theta, \pi_\theta) = G(0) > 0$ for any θ, and $\overline{\kappa}(\theta, \theta') = G(\theta - \theta')/G(0) =: \mathtt{K}(\theta - \theta') \leq 1$. □

Shift-invariant kernels are intimately connected with random Fourier features via Bochner's theorem [Rahimi and Recht, 2008]. For technical reasons, we consider weighted variants of random Fourier features. In fact, observe that the integral kernel given by (46) is invariant if we rescale the features by a weight function $w(\omega)^{-1}$ and their distribution Λ by $w^2(\omega)$. This additional freedom in the design of random features corresponding to a given kernel is convenient to obtain appropriate control of the moments of Λ involving powers of the frequency, as will be needed below.

Definition 6.3 (Weighted random Fourier features)**.** *Consider $\Omega = \mathcal{Z} = \mathbb{R}^d$, $w : \Omega \to \mathbb{R}$ a function such that $\inf_\omega w(\omega) = w(0) = 1$, and $\mathcal{F} = \{\phi_\omega\}_{\omega\in\Omega}$, with*

$$\phi_\omega(x) := \frac{e^{\jmath\langle\omega, x\rangle}}{w(\omega)} \qquad \forall x \in \mathbb{R}^d. \tag{76}$$

Given an arbitrary probability distribution Λ on the vector $\omega \in \mathbb{R}^d$, the corresponding kernel given by (46) *is shift invariant.*

6.2 Ingredients to apply Theorem 5.13

To exploit Theorem 5.13, a first ingredient is to control $\|\mathcal{D}\|_\mathcal{F}$ via Theorem 5.15 and a characterization of the quantities $\|\mathcal{M}\|_\mathcal{F}$ and $L_\mathcal{F}$.

Lemma 6.4. *Consider $\mathcal{T} = (\mathbf{\Theta}, \|\cdot\|, \varphi)$ a location family built from a probability distribution π_0 and denote $\|\cdot\|_\star$ the dual norm defined for $u \in \mathbb{R}^d$ by*

$$\|u\|_\star := \sup_{v:\|v\|\leq 1} u^T v. \tag{77}$$

Consider a shift-invariant kernel κ on $\mathbb{R}^d$ such that $\|\pi_0\|_\kappa > 0$, $\mathcal{M} = \mathcal{M}_\kappa(\mathcal{T})$, and $\psi : \theta \mapsto \pi_\theta/\|\pi_\theta\|_\kappa$.

Consider w a weight function and $\mathcal{F} = \{\phi_\omega\}_{\omega\in\Omega}$ as in Definition 6.3, and let $L_\mathcal{F}$ be the Lipschitz constant of ψ with respect to $\|\cdot\|$ and $\|\cdot\|_\mathcal{F}$. Denoting $\mathcal{F}' := \{\|\omega\|_\star \phi_\omega\}_{\omega\in\Omega}$, we have

$$\|\mathcal{M}\|_\mathcal{F} = \|\pi_0\|_\kappa^{-1} \cdot \|\pi_0\|_\mathcal{F} = \|\pi_0\|_\kappa^{-1};$$
$$L_\mathcal{F} = \|\pi_0\|_\kappa^{-1} \cdot \sup_\omega |\langle \pi_0, \|\omega\|_\star \phi_\omega\rangle| = \|\pi_0\|_\kappa^{-1} \cdot \|\pi_0\|_{\mathcal{F}'}.$$

Proof. For $\theta \in \mathbf{\Theta}$ and $\omega \in \mathbb{R}^d$ we have

$$\langle \pi_\theta, \phi_\omega\rangle = \mathbb{E}_{X\sim\pi_\theta} e^{\jmath\langle\omega,X\rangle}/w(\omega) = \mathbb{E}_{X\sim\pi_0} e^{\jmath\langle\omega,\theta+X\rangle}/w(\omega) = \langle \pi_0, \phi_\omega\rangle e^{\jmath\langle\omega,\theta\rangle} \tag{78}$$

hence $\|\pi_\theta\|_\mathcal{F} = \sup_\omega |\langle \pi_0, \phi_\omega\rangle| = \sup_\omega \left|\mathbb{E}_{X\sim\pi_0} e^{\jmath\langle\omega,X\rangle}/w(\omega)\right| \leq 1$ since $w(\omega) \geq 1$. The bound is achieved for $\omega = 0$ since $w(0) = 1$. Now, by definition, any $\mu \in \mathcal{M}$ can be written as $\mu = \psi(\theta) = \pi_\theta/\|\pi_\theta\|_\kappa$. Since κ is shift-invariant we have $\|\pi_\theta\|_\kappa = \|\pi_0\|_\kappa$ hence $\|\mu\|_\mathcal{F} = \|\pi_\theta\|_\mathcal{F}/\|\pi_\theta\|_\kappa = \|\pi_0\|_\kappa^{-1}$ is independent of θ and $\|\mathcal{M}\|_\mathcal{F} := \sup_{\mu\in\mathcal{M}} \|\mu\|_\mathcal{F} = \|\pi_0\|_\kappa^{-1}$. Another consequence of (78) is that $\langle \psi(\theta), \phi_\omega\rangle = \|\pi_0\|_\kappa^{-1} \langle \pi_0, \phi_\omega\rangle e^{\jmath\langle\omega,\theta\rangle}$. For $a \leq b$, $\left|e^{\jmath a} - e^{\jmath b}\right| = \left|\int_a^b \jmath e^{\jmath u} du\right| \leq \int_a^b |\jmath e^{\jmath u}| du = b - a$, hence

$$\begin{aligned}
\|\psi(\theta') - \psi(\theta)\|_\mathcal{F} &= \|\pi_0\|_\kappa^{-1} \cdot \sup_\omega \left\{ |\langle \pi_0, \phi_\omega\rangle| \cdot \left| e^{\jmath\langle\omega,\theta'\rangle} - e^{\jmath\langle\omega,\theta\rangle} \right| \right\} \leq \|\pi_0\|_\kappa^{-1} \cdot \sup_\omega \{ |\langle \pi_0, \phi_\omega\rangle| \cdot |\langle \omega, \theta' - \theta\rangle| \} \\
&\leq \|\pi_0\|_\kappa^{-1} \cdot \sup_\omega \{ |\langle \pi_0, \phi_\omega\rangle| \cdot \|\omega\|_\star \} \cdot \|\theta' - \theta\| = \|\pi_0\|_\kappa^{-1} \cdot \sup_\omega |\langle \pi_0, \|\omega\|_\star \phi_\omega\rangle| \cdot \|\theta' - \theta\| \\
&= \|\pi_0\|_\kappa^{-1} \cdot \|\pi_0\|_{\mathcal{F}'} \cdot \|\theta' - \theta\|.
\end{aligned}$$

To conclude we show that the bound is tight. When $\|\pi_0\|_{\mathcal{F}'}$ is finite (resp. infinite), for each integer $n \geq 1$ there is $\omega_n \neq 0$ such that $|\langle \pi_0, \|\omega_n\|_\star \phi_{\omega_n}\rangle| \geq \|\pi_0\|_{\mathcal{F}'} - 1/n$ (resp. $\geq n$). By compactness of the unit ball of $\|\cdot\|$ in $\mathbb{R}^d$ there is u_n such that $\|u_n\| = 1$ and $\langle \omega_n, u_n\rangle = \|\omega_n\|_\star$. Setting $\theta'_n = \frac{u_n}{n\|\omega_n\|_\star}$ and $\theta = 0$ we get $\langle \omega_n, \theta'_n\rangle = 1/n$ and $\langle \omega_n, \theta\rangle = 0$, so that $\left|e^{\jmath\langle\omega,\theta'_n\rangle} - e^{\jmath\langle\omega,\theta\rangle}\right| \overset{n\to\infty}{\sim} 1/n$. Straightforward arguments then show that $\lim_{n\to\infty} \|\psi(\theta'_n) - \psi(\theta)\|_\mathcal{F}/\|\theta'_n - \theta\| \geq \|\pi_0\|_\kappa^{-1} \cdot \|\pi_0\|_{\mathcal{F}'}$. □

In order to leverage Theorem 5.13, we now exhibit λ, γ such that (66) holds (Lemma 6.5 below), and more concrete estimates for the covering numbers $\mathrm{N}(\|\cdot\|_\mathcal{F}, \mathcal{D}, \delta)$ (Lemma 6.7 below) which are pivotal to determine the required number of measurements.

Lemma 6.5. *Consider $\mathcal{T} = (\mathbf{\Theta}, \|\cdot\|, \varphi)$ a location family built from a probability distribution π_0. Consider w a weight function, $\mathcal{F} = \{\phi_\omega\}_{\omega\in\Omega}$, Λ a probability distribution on Ω and κ the associated shift-invariant kernel as in Definition 6.3.*

Assume that $\|\pi_0\|_\kappa > 0$ and let $\overline{\kappa} : \mathbb{R}^d \to \mathbb{R}$ be the function associated to the $\mathcal{T}$-normalized version of κ as in Proposition 6.2. Assume there exists $a > 0, b \geq 1/2$ such that

$$1 - \overline{\kappa}(x) \geq \min(1, (\|x\|/a)^2)/b, \ \forall x \text{ s.t. } \|x\| \leq 1, \tag{79}$$

and $\lambda_0 > 0$ such that for each $u \in \mathbb{R}^d$ such that $\|u\| = 1$ and each integer $q \geq 2$ we have

$$\mathbb{E}_{\omega\sim\Lambda} \left\{ |\langle \pi_0, \phi_\omega\rangle|^{2q} \cdot \langle \omega, u\rangle^{2q} \right\} \leq \|\pi_0\|_\kappa^2 \tfrac{q!}{2} \lambda_0^q. \tag{80}$$

Then for each integer $q \geq 2$ we have

$$\sup_{\mu\in\mathcal{D}} \mathbb{E}_{\omega\sim\Lambda} |\langle \mu, \phi_\omega\rangle|^{2q} \leq \frac{q!}{2} [\lambda \|\pi_0\|_\kappa^{-2}]^{q-1} \cdot \lambda \qquad \text{with } \lambda := \max(2b, 1 + ba^2\lambda_0/2) \geq 1. \tag{81}$$

The proof is in Appendix C.1.

Remark 6.6. *If κ is c-strongly locally characteristic with respect to $\mathcal{T}$ with $0 < c \leq 2$ then (79) holds with $b = 2/c \geq 1$ and $a = 1$. With specific choices of π_0, Λ discussed in Section 6.3 we obtain finer estimates and provide concrete bounds for λ_0, $\|\pi_0\|_\kappa$, $\|\pi_0\|_{\mathcal{F}'}$, a and b.*

Lemma 6.7. *Let $\mathcal{T}$ be a location family based on a probability distribution π_0 on $\mathbb{R}^d$ and a norm $\|\cdot\|$. Assume that the covering numbers of the base parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ satisfy, for some $C_{\mathcal{T}} \geq 1$,*

$$\mathrm{N}(\|\cdot\|, \boldsymbol{\Theta}, \delta) \leq \max\bigl(1, \tfrac{C_{\mathcal{T}}}{\delta}\bigr)^d, \qquad \delta > 0. \tag{82}$$

Consider $\mathcal{F}$ a weighted random Fourier feature family as in Definition 6.3, κ the induced shift-invariant kernel, and $\mathcal{D} = \mathcal{D}_\kappa(\mathcal{T})$ the induced set of normalized dipoles. Denote $\mathcal{F}' := \{\|\omega\|_\star \phi_\omega\}_{\omega\in\Omega}$ and $\mathcal{F}'' := \left\{\|\omega\|_\star^2 \phi_\omega\right\}_{\omega\in\Omega}$. If κ is 1-strongly[8] locally characteristic on $\mathcal{T}$ then, defining $D := \|\mathcal{D}\|_{\mathcal{F}} \geq 1$ and

$$\begin{aligned}
C_{\mathcal{F}} &= \|\pi_0\|_\kappa^{-1} \|\pi_0\|_{\mathcal{F}} = \|\pi_0\|_\kappa^{-1};\\
C'_{\mathcal{F}} &= \|\pi_0\|_\kappa^{-1} \|\pi_0\|_{\mathcal{F}'} = \|\pi_0\|_\kappa^{-1} \sup_\omega \{|\langle \pi_0, \phi_\omega\rangle| \|\omega\|_\star\};\\
C''_{\mathcal{F}} &= \|\pi_0\|_\kappa^{-1} \|\pi_0\|_{\mathcal{F}''} = \|\pi_0\|_\kappa^{-1} \sup_\omega \left\{|\langle \pi_0, \phi_\omega\rangle| \|\omega\|_\star^2\right\}
\end{aligned}$$

it holds

$$\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \delta) \leq 2 \max\Bigl(1, \tfrac{64 C_{\mathcal{T}} (D C''_{\mathcal{F}} + C'_{\mathcal{F}} + C_{\mathcal{F}})}{\delta}\Bigr)^{4(d+1)}, \qquad \delta > 0. \tag{83}$$

The proof is in Appendix C.2.

6.3 Random Fourier sketching with a Gaussian kernel

For the two scenarios of Sections 3-4, clustering and compressive GMM, the natural model set $\mathfrak{S}^\star(\mathcal{H})$ is location-based, either built with Diracs ($\pi_0 = \delta_0$) or Gaussians ($\pi_0 = \mathcal{N}(0, \boldsymbol{\Sigma})$). For these scenarios, the following distribution Λ of random frequencies is specifically designed to lead to a Gaussian kernel when matched with weighted random Fourier features (Definition 6.3) using the same weight function.

Definition 6.8 (Frequency distribution)**.** *Let $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$ be positive definite, and denote $p_{\mathcal{N}(0,\boldsymbol{\Gamma})}(\omega)$ the probability density function (pdf) of the centered Gaussian with covariance $\boldsymbol{\Gamma}$. Given a weight function w as in Definition 6.3, define a probability distribution Λ on the frequency ω through the pdf*

$$\Lambda(\omega) := C_\Lambda^{-2} w^2(\omega)\ p_{\mathcal{N}(0,\boldsymbol{\Gamma})}(\omega)\,, \tag{84}$$

where

$$C_\Lambda := \sqrt{\mathbb{E}_{\omega\sim\mathcal{N}(0,\boldsymbol{\Gamma})} w^2(\omega)}, \tag{85}$$

Since $\inf_\omega w(\omega) = 1$ we have $C_\Lambda \geq 1$. When using the unit weight function $w \equiv 1$ we get $C_\Lambda = 1$.

From Definitions 6.3 and 6.8 one can build a random feature map Φ as in Section 5.1. Its properties depend on the choice of the weight function w (which will always be chosen identical in the definition

[8] the result is easily adjusted if κ is c-strongly locally characteristic with $c < 1$.

of $\mathcal{F}$ and Λ) and of the covariance matrix $\mathbf{\Gamma}$. Before discussing the choice of these parameters, one can immediately observe that the associated kernel is Gaussian: for any $x, x' \in \mathcal{Z}$ we have

$$\begin{aligned}\kappa(x,x') &= \mathbb{E}_{\omega\sim\Lambda}\phi_\omega(x)\overline{\phi_\omega(x')} = \int_{\omega\in\mathbb{R}^d} w^{-2}(\omega)e^{\jmath\omega^T(x-x')}C_\Lambda^{-2}w^2(\omega)p_{\mathcal{N}(0,\mathbf{\Gamma})}(\omega)d\omega \\ &= C_\Lambda^{-2}\cdot\mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{\Gamma})}e^{\jmath\omega^T(x-x')} \overset{(*)}{=} C_\Lambda^{-2}\cdot\exp\left(-\frac{\|x-x'\|_{\mathbf{\Gamma}^{-1}}^2}{2}\right), \end{aligned} \tag{86}$$

where (*) follows from the expression of the characteristic function of the Gaussian and $\|c\|_{\mathbf{A}} := \sqrt{c^T\mathbf{A}^{-1}c}$ is the Mahalanobis norm (36) given a positive definite matrix $\mathbf{A}$. We focus on two scenarios.

Definition 6.9. *Consider $\varepsilon > 0$ some separation, $s > 0$ some scale, and $\mathbf{\Theta} \subseteq \mathbb{R}^d$ some parameter space. We consider the following setting using Definitions 6.1,6.3,6.8.*

- **for mixtures of Diracs***: $\mathcal{T}_{\texttt{Dirac}}$ is defined with $\pi_0 = \delta_0$, and $\|\cdot\| = \|\cdot\|_2/\varepsilon$; $\mathcal{F}_{\texttt{Dirac}}$ is defined with a weight function $w(\cdot)$ to be discussed; $\Lambda_{\texttt{Dirac}}$ is defined with the same $w(\cdot)$ and $\mathbf{\Gamma} = s^{-2}\mathbf{I}_d$. In several places we focus more specifically on $\mathbf{\Theta} = \mathbf{\Theta}_R := \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,R)$, where $R \geq \varepsilon$.*
- **for mixtures of Gaussians***: define $\mathcal{T}_{\texttt{Gauss}}$ with $\pi_0 = \mathcal{N}(0,\mathbf{\Sigma})$ for some chosen positive definite $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ and $\|\cdot\| = \|\cdot\|_{\mathbf{\Sigma}}/\varepsilon$; $\mathcal{F}_{\texttt{Gauss}}$ is defined with $w(\cdot) \equiv 1$; $\Lambda_{\texttt{Gauss}}$ is defined with the same $w(\cdot)$ and $\mathbf{\Gamma} = s^{-2}\mathbf{\Sigma}^{-1}$. Again for some results we will focus on $\mathbf{\Theta}_R := \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\mathbf{\Sigma}}}(0,R)$, where $R \geq \varepsilon$.*

Observe that in both cases we have the identity $\mathbf{\Theta}_R = \mathcal{B}_{\mathbb{R}^d,\|\cdot\|}(0,R/\varepsilon)$, and that $\|\theta-\theta'\| = \|\psi(\theta)-\psi(\theta')\|_2$ with $\psi_{\texttt{Dirac}}(\theta) = \theta/\varepsilon$, while $\psi_{\texttt{Gauss}}(\theta) = \mathbf{\Sigma}^{-1/2}\theta/\varepsilon$.

6.3.1 Properties of the kernel mean embedding

Before we state the main theorem of this section, we make a few observations. For the Dirac scenario, since $\mathbf{\Gamma} = s^{-2}\mathbf{I}_d$, the kernel mean embedding associated to (86) satisfies

$$\kappa(\pi_\theta,\pi_{\theta'}) = \kappa(\theta,\theta') = C_\Lambda^{-2}\cdot\exp\left(-\tfrac{1}{2}\cdot\frac{\|\theta-\theta'\|_2^2}{s^2}\right) = C_\Lambda^{-2}\cdot\exp\left(-\tfrac{1}{2}\cdot\frac{\|\theta-\theta'\|^2}{(s/\varepsilon)^2}\right)$$

For Gaussians, as $w \equiv 1$ we have $C_\Lambda = 1$. By Lemma C.3 (see Appendix C.3) with $\mathbf{R} = \mathbf{\Gamma}^{-1} = s^2\mathbf{\Sigma}$ we obtain

$$\begin{aligned}\kappa(\pi_\theta,\pi_{\theta'}) =& \frac{\sqrt{\det(s^2\mathbf{\Sigma})}}{\sqrt{\det((2+s^2)\mathbf{\Sigma})}}\exp\left(-\tfrac{1}{2}\|\theta-\theta'\|_{(2+s^2)\mathbf{\Sigma}}^2\right) \\ =& \left(\frac{s^2}{2+s^2}\right)^{d/2}\exp\left(-\tfrac{1}{2}\cdot\frac{\|\theta-\theta'\|_{\mathbf{\Sigma}}^2}{2+s^2}\right) = \left(\frac{1}{1+2/s^2}\right)^{d/2}\exp\left(-\tfrac{1}{2}\cdot\frac{\|\theta-\theta'\|^2}{(2+s^2)/\varepsilon^2}\right).\end{aligned}$$

This yields

$$\|\pi_0\|_\kappa = \begin{cases} C_\Lambda^{-1} = \left[\mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)}w^2(\omega)\right]^{-1/2}, & \text{for Diracs,} \\ \left(1+2/s^2\right)^{-d/4}, & \text{for Gaussians.}\end{cases} \tag{87}$$

In both cases we have $\|\pi_0\|_\kappa \leq 1$, and we also get

$$\overline{\kappa}(\theta-\theta') := \tfrac{\kappa(\pi_\theta,\pi_{\theta'})}{\|\pi_\theta\|_\kappa\|\pi_{\theta'}\|_\kappa} = K_{\sigma(s)/\varepsilon}(\|\theta-\theta'\|), \quad \forall\theta,\theta' \in \mathbb{R}^d, \tag{88}$$

with

$$K_\sigma(u) := e^{-\frac{u^2}{2\sigma^2}}, \quad u \geq 0; \qquad \text{and} \qquad \sigma(s) := \begin{cases} s, & \text{for Diracs} \\ \sqrt{2+s^2}, & \text{for Gaussians.}\end{cases} \tag{89}$$

The following properties of Gaussian kernels are proved in Appendix C.4. The first property allows to use Lemma 6.5. The second shows that the kernel is 1-strongly locally characteristic. The third one shows that its $2k$-coherence is below $3/4$ provided $\sigma \lesssim 1/\sqrt{\log k}$.

Lemma 6.10. *Consider $\sigma > 0$.*

1. *We have $1 - K_\sigma(u) \geq \min(1, (u/\sigma)^2)/3$ for all $u \geq 0$;*
2. *If $\sigma \leq 1/\sqrt{2}$ then $0 \leq K_\sigma(u) \leq 1 - u^2/2$ for $u \in [0, 1]$;*
3. *Given an integer $k \geq 1$, for any $\sigma \leq \sigma_k^\star := (4\sqrt{\log(ek)})^{-1}$, it holds $C(K_\sigma) \leq \frac{3}{16(2k-1)}$, where $C(K)$ is defined in (74).*

Using the generic tools of Section 5 with the model set $\mathfrak{S}_k(\mathcal{T})$ we can establish the following result.

Theorem 6.11. *Consider $\mathcal{F}_{\texttt{Dirac}}, \Lambda_{\texttt{Dirac}}, \mathcal{T}_{\texttt{Dirac}}$ (resp. $\mathcal{F}_{\texttt{Gauss}}, \Lambda_{\texttt{Gauss}}, \mathcal{T}_{\texttt{Gauss}}$) as in Definition 6.9 with separation ε, scale s and weight w such that*

$$\varepsilon = \begin{cases} s/\sigma_k^\star, & \text{for Diracs;} \\ \sqrt{2+s^2}/\sigma_k^\star, & \text{for Gaussians.} \end{cases} \tag{90}$$

where $k \geq 1$ is an integer and $\sigma_k^\star := (4\sqrt{\log(ek)})^{-1}$. Then, the associated kernel κ is 1-strongly characteristic and has its $2k$-coherence bounded by $3/4$.
Further assume that there exists[9] *$C_{\boldsymbol{\Theta}} \geq \varepsilon$ such that the parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ satisfies*

$$\max\big(1, \tfrac{C_{\boldsymbol{\Theta}}}{\delta}\big)^d \geq \begin{cases} \mathrm{N}(\|\cdot\|_2, \boldsymbol{\Theta}, \delta), & \text{for Diracs,} \\ \mathrm{N}(\|\cdot\|_{\boldsymbol{\Sigma}}, \boldsymbol{\Theta}, \delta), & \text{for Gaussians,} \end{cases}, \qquad \delta > 0, \tag{91}$$

and denote

$$A := \begin{cases} \mathbb{E}_{\omega \sim \mathcal{N}(0, s^{-2}\mathbf{I}_d)} w^2(\omega), & \text{for Diracs,} \\ (1 + 2/s^2)^{d/2}, & \text{for Gaussians,} \end{cases} \quad \text{and } B := \begin{cases} 1 + \varepsilon^2 \Big(\sup_\omega \frac{\|\omega\|_2}{w(\omega)}\Big)^2, & \text{for Diracs,} \\ 1 + \varepsilon^2, & \text{for Gaussians;} \end{cases} \tag{92}$$

$$C := \begin{cases} 64A\sqrt{2B}C_{\boldsymbol{\Theta}}\varepsilon^{-1}\Big(1 + \varepsilon \sup_\omega \frac{\|\omega\|_2}{w(\omega)} + \varepsilon^2 \sup_\omega \frac{\|\omega\|_2^2}{w(\omega)}\Big), & \text{for Diracs,} \\ 64A\sqrt{2B}C_{\boldsymbol{\Theta}}\varepsilon^{-1}(1 + \varepsilon + \varepsilon^2), & \text{for Gaussians.} \end{cases} \tag{93}$$

Consider $0 < \delta, \zeta < 1$ and $\Phi(x) := \frac{1}{\sqrt{m}}\big(\phi_{\omega_j}(x)\big)_{j=1}^m$, where $(\omega_j)_{j=1}^m$ are drawn i.i.d. according to Λ. If

$$m \geq 80 \cdot \delta^{-2} \cdot A \cdot \min(12e\log^2(24ek), 2B) \cdot k \cdot \{2k(4d+5)[\log(kCAB) + \log(2048/\delta)] + \log(2/\zeta)\}, \tag{94}$$

then with probability at least $1 - \zeta$ on the draw of $(\omega_j)_{j=1}^m$ the operator $\mathcal{A}$ induced by Φ satisfies

$$1 - \delta \leq \frac{\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} \leq 1 + \delta, \qquad \forall \pi, \pi' \in \mathfrak{S}_k(\mathcal{T}). \tag{95}$$

When (95) holds, the LRIP (11) holds with $C_{\mathcal{A}} := \frac{8\sqrt{2k}\|\mathcal{D}\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$ and $\eta = 0$ for each loss class $\mathcal{L}$.

Remark 6.12. *In light of the separation assumption (90) for Gaussians, which implies $\varepsilon \geq \sqrt{2}/\sigma_k^\star = 4\sqrt{2\log(ek)}$, the dominant term of the rightmost factor in C (cf (93)) is ε^2 for Gaussians. In contrast, for Diracs, the rightmost factor in C can become arbitrarily close to 1 by setting ε (and s) small enough.*

[9] If needed, consider $C'_{\boldsymbol{\Theta}} = \max(C_{\boldsymbol{\Theta}}, \varepsilon)$.

Proof. The proof consists in checking the assumptions of the generic Theorem 5.13 using the more concrete estimates obtained previously in this section in the mixture of Dirac and Gaussian settings.

Step 1: control of the coherence. Assumption (90) means that $\sigma(s)/\varepsilon{=}\sigma_k^\star$ with $\sigma(s)$ as in (89). By Lemma 6.10 it follows that $K_{\sigma(s)/\varepsilon}$ satisfies $0 \le K_{\sigma(s)/\varepsilon}(u) \le 1 - u^2/2$ for $0 \le u \le 1$ and that $(2k-1)C(K_{\sigma/\varepsilon}) \le 3/16$ with $C(K)$ defined in (74). Since $\|\theta - \theta'\| = \|\psi(\theta) - \psi(\theta')\|_2$ with $\psi(\theta) = \theta/\varepsilon$ for Diracs (resp. $\psi(\theta) := \mathbf{\Sigma}^{-1/2}\theta/\varepsilon$ for Gaussians), by property (88) and Theorem 5.16 we obtain that κ is 1-strongly locally characteristic with respect to $\mathcal{T}$, and that κ has coherence M with respect to $\mathcal{T}$, where $(2k-1)M \le (2k-1)4C(K_{\sigma/\varepsilon}) \le 3/4$.

Step 2: control of $\|\mathcal{D}\|_{\mathcal{F}}, \|\pi_0\|_\kappa, \|\pi_0\|_{\mathcal{G}}, \mathcal{G} \in \{\mathcal{F}, \mathcal{F}', \mathcal{F}''\}$. In light of property (87), assumption (92) and Lemma 6.4 we have $\|\mathcal{M}\|_{\mathcal{F}}^2 = \|\pi_0\|_{\mathcal{F}}^2 \cdot \|\pi_0\|_\kappa^{-2} = \|\pi_0\|_\kappa^{-2} = A$. To control $L_{\mathcal{F}} = C'_{\mathcal{F}}$ and $C''_{\mathcal{F}}$ as in Lemmas 6.4 and 6.7 we compute:

- for Diracs: as $\|\cdot\| = \|\cdot\|_2/\varepsilon$ we have $\|\cdot\|_\star = \varepsilon\|\cdot\|_2$, and since $\pi_0 = \delta_0$ we get

$$\|\pi_0\|_{\mathcal{F}'} := \sup_\omega |\langle \pi_0, \|\omega\|_\star \phi_\omega\rangle| = \sup_\omega \|\omega\|_\star |\phi_\omega(0)| = \varepsilon \sup_\omega \frac{\|\omega\|_2}{w(\omega)};$$
$$\|\pi_0\|_{\mathcal{F}''} := \sup_\omega \left|\left\langle \pi_0, \|\omega\|_\star^2 \phi_\omega \right\rangle\right| = \sup_\omega \|\omega\|_\star^2 |\phi_\omega(0)| = \varepsilon^2 \sup_\omega \frac{\|\omega\|_2^2}{w(\omega)};$$

- for Gaussians: since $\|\cdot\| = \|\cdot\|_{\mathbf{\Sigma}}/\varepsilon$ we have $\|\cdot\|_\star = \varepsilon\|\cdot\|_{\mathbf{\Sigma}^{-1}}$, and since $w(\omega) \equiv 1$, by the expression of the characteristic function of Gaussians we have $|\langle \pi_0, \phi_\omega\rangle| = e^{-\|\omega\|_{\mathbf{\Sigma}^{-1}}^2/2}$, hence

$$\|\pi_0\|_{\mathcal{F}'} := \sup_\omega |\langle \pi_0, \|\omega\|_\star \phi_\omega\rangle| = \varepsilon \sup_\omega \|\omega\|_{\mathbf{\Sigma}^{-1}} e^{-\|\omega\|_{\mathbf{\Sigma}^{-1}}^2/2} = \varepsilon \sup_{u \ge 0} u e^{-u^2/2} = \varepsilon e^{-1/2} \le \varepsilon;$$
$$\|\pi_0\|_{\mathcal{F}''} := \sup_\omega \left|\left\langle \pi_0, \|\omega\|_\star^2 \phi_\omega \right\rangle\right| = \varepsilon^2 \sup_\omega \|\omega\|_{\mathbf{\Sigma}^{-1}}^2 e^{-\|\omega\|_{\mathbf{\Sigma}^{-1}}^2/2} = \varepsilon^2 \sup_{u \ge 0} u e^{-u/2} = \varepsilon^2 \cdot 2/e \le \varepsilon^2.$$

Since κ is 1-strongly locally characteristic, by Theorem 5.15, Lemma 6.4, (92) and $(a+b)^2 \le 2(a^2+b^2)$ it follows that

$$\|\mathcal{D}\|_{\mathcal{F}}^2 \le 2\|\mathcal{M}\|_{\mathcal{F}}^2 + 2L_{\mathcal{F}}^2 {=} 2A(1 + \|\pi_0\|_{\mathcal{F}'}^2) = 2AB.$$

Step 3: control of normalized dipole moments. We show in Appendix C.5 that in both settings, for each $u \in \mathbb{R}^d$ such that $\|u\| = 1$ and each integer $q \ge 2$, we have

$$\mathbb{E}_{\omega\sim\Lambda}\left\{ |\langle \pi_0, \phi_\omega\rangle|^{2q} \langle \omega, u\rangle^{2q} \right\} \quad \le \|\pi_0\|_\kappa^2 (2\varepsilon^2/\sigma^2(s))^q \tfrac{q!}{2}. \tag{96}$$

This proves (80) with $\lambda_0 = 2\varepsilon^2/\sigma^2(s) = 2/(\sigma_k^\star)^2$. By Lemma 6.10-1 applied to $K_{\sigma(s)/\varepsilon}$ and property (88) we obtain that $\overline{\kappa}(x) = K_{\sigma(s)/\varepsilon}(\|x\|)$ satisfies (79) with $a = \sigma(s)/\varepsilon = \sigma_k^\star$ and $b = 3$. By Lemma 6.5, since $1 + ba^2\lambda_0/2 = 4$ while $2b = 6$, we get that

$$\sup_{\mu\in\mathcal{D}} \mathbb{E}_{\omega\sim\Lambda} |\langle \mu, \phi_\omega\rangle|^{2q} \le \frac{q!}{2} [6\|\pi_0\|_\kappa^{-2}]^{q-1} \cdot 6.$$

It follows that the assumption (66) of Theorem 5.13 holds with $\gamma = 6\|\pi_0\|_\kappa^{-2} = 6A$, $\lambda = 6$.

Remark 6.13. *Since κ is 1-strongly locally characteristic (cf Lemma 6.10-2), the generic arguments in Remark 6.6 show that $\overline{\kappa}$ also satisfies (79) with $a' = 1$ and $b' = 2$. Notice that here since $\sigma(s)/\varepsilon{=}\sigma_k^\star$ we have $ba^2 = 3(\sigma_k^\star)^2 = \mathcal{O}(1/\log(ek)) \ll 2 = b'(a')^2$ for large k, showing that Lemma 6.10-1 indeed allows to improve over the generic result of Remark 6.6.*

Step 4: covering numbers for $\mathcal{D}$. To control those covering numbers, we apply Lemma 6.7. The final estimate (83) from this lemma involves the quantity $64C_{\mathcal{T}}(C_{\mathcal{F}} + C'_{\mathcal{F}} + \|\mathcal{D}\|_{\mathcal{F}} C''_{\mathcal{F}})$, which we now show is bounded by C defined in (93) in both cases, using the estimates from Step 2 for the various constants. Since $\|\cdot\| = \|\cdot\|_2/\varepsilon$ (for Diracs), resp. $\|\cdot\| = \|\cdot\|_{\boldsymbol{\Sigma}}/\varepsilon$ (for Gaussians), Assumption (91) implies that Assumption (82) of Lemma 6.7 holds with $C_{\mathcal{T}} := \varepsilon^{-1} C_{\boldsymbol{\Theta}}$.

For Diracs: From the estimates in Step 2 and the fact that $A \geq 1, B \geq 1$ we obtain

$$\begin{aligned} 64C_{\mathcal{T}}(C_{\mathcal{F}} + C'_{\mathcal{F}} + \|\mathcal{D}\|_{\mathcal{F}} C''_{\mathcal{F}}) &= 64C_{\boldsymbol{\Theta}}\varepsilon^{-1}\|\pi_0\|_\kappa^{-1}(\|\pi_0\|_{\mathcal{F}} + \|\pi_0\|_{\mathcal{F}'} + \|\mathcal{D}\|_{\mathcal{F}}\|\pi_0\|_{\mathcal{F}''}) \\ &\leq 64C_{\boldsymbol{\Theta}}\varepsilon^{-1}\sqrt{A}\left(1 + \varepsilon \sup_\omega \tfrac{\|\omega\|_2}{w(\omega)} + \sqrt{2AB}\varepsilon^2 \sup_\omega \tfrac{\|\omega\|_2^2}{w(\omega)}\right) \leq C. \end{aligned}$$

For Gaussians: From the estimates in Step 2 we obtain similarly to the Dirac case

$$64C_{\mathcal{T}}(C_{\mathcal{F}} + C'_{\mathcal{F}} + \|\mathcal{D}\|_{\mathcal{F}} C''_{\mathcal{F}}) \leq 64C_{\boldsymbol{\Theta}}\varepsilon^{-1}A\sqrt{2B}(1 + \varepsilon + \varepsilon^2) = C.$$

By Lemma 6.7 we obtain $\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \delta) \leq 2(C/\delta)^r$ for each $0 < \delta < 1$, with $r = 4(d+1)$. Since $C \geq 1$ ($A, B, C_{\boldsymbol{\Theta}}\varepsilon^{-1}$ being greater than 1), the covering number assumption of Theorem 5.13 is satisfied.

Step 5: wrapping up. Combining the above ingredients we can apply Theorem 5.13. We have

$$\min\left(2e\gamma \log^2(4ek\lambda), \|\mathcal{D}\|_{\mathcal{F}}^2\right) \leq A \min(12e \log^2(24ek), 2B);$$

combined with (94), the above estimates imply that (67) holds. We conclude using Theorem 5.13. □

6.4 Additional ingredients to establish Theorem 3.1 and Theorem 4.1

The detailed proofs of Theorem 3.1 (resp. of Theorem 4.1) are given in Appendix D. The proofs use as an intermediate tool a constrained hypothesis class $\mathcal{H}$ such that the model set $\mathfrak{S}^\star(\mathcal{H})$ (resp. $\mathfrak{S}^{\mathtt{ML}}(\mathcal{H})$) corresponds to k-separated mixtures of Diracs (resp. of Gaussians). As both compressive clustering (see Section 3) and compressive Gaussian Mixture Modeling (see Section 4) rely on (reweighted) random Fourier features to design the sketching function Φ, this allows to leverage Theorem 6.11 combined with Theorem 2.2 to establish intermediate results.

Additional ingredients to complete the proofs include:

- the constant $\|\mathcal{D}\|_{\Delta\mathcal{L}(\mathcal{H})}$ determining the LRIP constant $C_{\mathcal{A}}$ established as the product of Theorem 6.11. This constant is controlled in Appendix D.2;
- the bias term from Theorem 2.2, which is bounded by a more explicit estimate in Appendix D.3;

The theorems are proved in Appendix D.4.

7 Conclusion and perspectives

The principle of compressive statistical learning is to learn from large-scale collections by first summarizing the collection into a sketch vector made of empirical (random) moments, before solving a nonlinear least squares problem. The main contribution of this paper is to demonstrate on two examples (compressive clustering and compressive Gaussian mixture estimation –with fixed known covariance) that the excess risk of this procedure can be controlled, as well as the sketch size.

Sharpened estimates? Our demonstration of the validity of the compressive statistical learning framework for certain tasks is, in a sense, qualitative, and we expect that many bounds and constants are sub-optimal. This is the case for example of the estimated sketch sizes for which statistical learning guarantees have been established, and an immediate theoretical challenge is to sharpen these guarantees to match the empirical phase transitions observed empirically for compressive clustering and compressive GMM [Keriven et al., 2018, 2017]. For mixture models, as our proof technique involves Geshgorin's disc theorem, it is natural to wonder to what extent the involved constants can be tightened to get closer to sharp oracle inequalities, possibly at the price of larger sketch sizes. Overall, an important question to benchmark the quality of the established bounds (on achievable sketch sizes, on the separation assumptions used for k-mixtures, etc.) is of course to investigate corresponding lower-bounds.

Provably-good algorithms of bounded complexity? As the control of the excess risk relies on the (approximate) minimizer of a nonlinear least-squares problem (13), the results in this paper are essentially information-theoretic. Can we go beyond the heuristic optimization algorithms derived for compressive k-means and compressive GMM [Keriven et al., 2018, 2017] and characterize provably good, computationally efficient algorithms to obtain this minimizer ?

Promising directions revolve around recent advances in super-resolution imaging and low-rank matrix recovery. For compressive clustering (resp. compressive GMM), the similarity between the minimization of (4) (resp. (5)) and super-resolution imaging suggests to explore TV-norm minimization –a *convex* problem– techniques [Candès and Fernandez-Granda, 2013, De Castro et al., 2016, Duval and Peyré, 2015] and to seek generalized RIP guarantees [Traonmilin and Gribonval, 2018]. Further, to circumvent the difficulties of optimization (convex or not) in the space of finite signed measures, it may also be possible to adapt the recent guarantees obtained for certain nonconvex problems that directly leverage a convex "lifted" problem [see e.g. Li and Tang, 2017] without incurring the cost of actually computing in the lifted domain.

Finally, the computational cost of sketching itself can be further controlled [Chatalic et al., 2018] by replacing random Gaussian weights where possible with fast approximations [Le et al., 2013, Choromanski and Sindhwani, 2016, Bojarski et al., 2017]. This results in accelerations of the learning stage wherever matrix multiplications are exploited. To conduct the theoretical analysis of the resulting sketching procedure, one will need to analyze the kernels associated to these fast approximations.

Acknowledgements

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906), the german DFG (SFB-1294 "Data Assimilation"), the Franco-German University through the binational Doktorandenkolleg CDFA 01-18, and the ANR (ANR-19-CHIA-0021-01, project BISCOTTE; ANR-19-CHIA-0009, project AllegroAssai). Rémi Gribonval is very grateful to Michael E. Davies for many enlightening discussions around the idea of compressive statistical learning since this project started several years ago. The authors also wish to warmly thank Bernard Delyon and Adrien Saumard, as well as Gabriel Peyré and Lorenzo Rosasco for their constructive feedback on early versions of this manuscript.

Appendix

We begin by introducing notations and useful results. We then provide general properties on covering numbers, followed by properties that are shared by any model of mixtures of distributions that are sufficiently separated $\mathfrak{S} = \mathfrak{S}_k(\mathcal{T})$. We then apply these results to mixtures of Diracs and both k-medians and k-means risks, and to Gaussian Mixture Models with fixed known covariance for maximum likelihood estimation.

A Generalities on covering numbers

In this section we formulate generic results on covering numbers.

A.1 Basic properties

The definition used in this paper is that of *internal* covering numbers, meaning that the centers z_i of the covering balls are required to be included in the set Y being covered. Somehow counter-intuitively these covering numbers (for a fixed radius δ) are not necessarily increasing with the inclusion of sets: for instance, consider a set A formed by two points, included in set B which is a ball of radius δ. Suppose those two points diametrically opposed in B. We have $A \subset B$, but two balls of radius δ are required to cover A (since their centers have to be in A), while only one such ball is sufficient to cover B. Yet, as shown by the following lemma, the covering numbers of included sets still behave in a controlled manner.

Lemma A.1. *Let $A \subseteq B \subseteq X$ be subsets of a pseudometric set (X, d), and $\delta > 0$. Then,*

$$\mathrm{N}(d, A, \delta) \leq \mathrm{N}(d, B, \delta/2). \tag{97}$$

Proof. Let $(b_i)_{1 \leq i \leq N}$ be a $\delta/2$-covering of B. We construct a δ-covering $(a_i)_{i \in I}$ of A of cardinality at most N in the following way. For each $i = 1, \ldots, N$, consider $C_i := \mathcal{B}_{X,d}(b_i, \delta/2) \cap A$. If $C_i \neq \emptyset$, we replace b_i by an arbitrary point $a_i \in C_i$, otherwise we discard b_i. Note that in the first case, by the triangle inequality $C_i \subset \mathcal{B}_{X,d}(b_i, \delta/2) \subset \mathcal{B}_{X,d}(a_i, \delta)$. On the other hand, by the covering property, $\bigcup_{1 \leq i \leq N} C_i = A$. Therefore the set of a_is is a δ-covering of A. □

Lemma A.2. *Let (X, d) and (X', d') be two pseudometric sets, and $Y \subseteq X$, $Y' \subseteq X'$. If there exists a surjective function $f : Y \to Y'$ which is L-Lipschitz with $L > 0$, i.e. such that*

$$\forall x, y \in Y,\ d'(f(x), f(y)) \leq L d(x, y),$$

then for all $\delta > 0$ we have

$$\mathrm{N}(d', Y', \delta) \leq \mathrm{N}(d, Y, \delta/L). \tag{98}$$

Proof. Define $\delta_2 = \delta/L$, denote $N = \mathrm{N}(d, Y, \delta_2)$, and let $y_i \in Y$, $i = 1, ..., N$ be a δ_2-covering of Y. Consider $y' \in Y'$. There exists $y \in Y$ such that $f(y) = y'$ since f is surjective. For some $1 \leq i \leq N$ we have $d(y, y_i) \leq \delta_2$, hence we have

$$d'(y', f(y_i)) = d'(f(y), f(y_i)) \leq L d(y, y_i) \leq L\delta_2 = \delta.$$

Thus $\{f(y_i)\}_{i=1,\ldots,N}$ is a δ-covering of Y', and we have $\mathrm{N}(d', Y', \delta) \leq N$. □

Lemma A.3. *Let Y, Z be two subsets of a pseudometric set (X, d) and $\epsilon \geq 0$ such that the following holds:*

$$\forall z \in Z,\ \exists y \in Y,\ d(z, y) \leq \epsilon. \tag{99}$$

Then for all $\delta > 0$

$$\mathrm{N}(d, Z, 2(\delta + \epsilon)) \leq \mathrm{N}(d, Y, \delta). \tag{100}$$

Proof. Denote $N = \mathrm{N}(d, Y, \delta)$ and let $y_1, ..., y_N \in Y$ be a δ-covering of Y. For all $z \in Z$, by the assumption (99) there is $y \in Y$ such that $d(z, y) \leq \epsilon$, and subsequently there is an index i such that $d(z, y_i) \leq d(z, y) + d(y, y_i) \leq \delta + \epsilon$. This implies $Z \subset \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \epsilon)$, hence by Lemma A.1

$$\mathrm{N}(d, Z, 2(\delta + \epsilon)) \leq \mathrm{N}\left(d, \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \epsilon), \delta + \epsilon\right) \leq N.$$

□

Lemma A.4 (Cucker and Smale [2002], Prop. 5)**.** *Let $(X, \|\cdot\|)$ be a Banach space of finite dimension d. Then for any $x \in X$ and $R > 0$ we have for any $\delta > 0$*

$$\mathrm{N}\big(\|\cdot\|, \mathcal{B}_{X,\|\cdot\|}(x, R), \delta\big) \leq \max\left(1, \left(\frac{4R}{\delta}\right)^d\right) \tag{101}$$

NB: The result in [Cucker and Smale, 2002, Prop. 5] does not include $\max(1, \cdot)$. This obviously cannot hold for $\delta > 4R$ since the left hand side is at least one. The proof of [Cucker and Smale, 2002, Prop. 5] yields the result stated here.

A.2 "Clipped" Secant set

To control the covering numbers of the normalized secant set (50), or those of the normalized dipole set $\mathcal{D}$ (56) it will be convenient to control those of certain subsets of its normalized secant set.

Lemma A.5. *Consider X a vector space, $\|\cdot\|_a, \|\cdot\|_b : X \to [0, +\infty]$ two semi-norms, $X_a, X_b \subseteq X$ the subspaces where they are finite, and $Y \subseteq X_a$. Consider $\mathcal{Q} \subseteq Y^2$ and assume that for some constants $0 < A \leq B < \infty$,*

$$\forall (y, y') \in \mathcal{Q},\ A\|y - y'\|_b \leq \|y - y'\|_a \leq B\|y - y'\|_b < \infty. \tag{102}$$

Given $\eta > 0$, consider the following subset of the normalized secant of Y:

$$\mathcal{S}_\eta := \left\{ \frac{y - y'}{\|y - y'\|_b} \;\middle|\; (y, y') \in \mathcal{Q} \subset Y^2,\ \|y - y'\|_b > \eta \right\}.$$

For each $\delta > 0$ we have

$$\mathrm{N}(\|\cdot\|_a, \mathcal{S}_\eta, \delta) \leq \mathrm{N}^2\left(\|\cdot\|_a, Y, \frac{\delta\eta}{4(1 + B/A)}\right). \tag{103}$$

Proof. Define the (semi)norm on Y^2: $\|(y_1, y_2) - (y_1', y_2')\|_a = \|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a$ and note that we have trivially $\mathrm{N}\big(\|\cdot\|_a, Y^2, \delta\big) \leq \mathrm{N}^2(\|\cdot\|_a, Y, \delta/2)$. Consider the set $\mathcal{Q}' := \{(y_1, y_2) \in \mathcal{Q} : \|y_1 - y_2\|_b > \eta\}$. By definition the function $f : (\mathcal{Q}', \|\cdot\|_a) \to (\mathcal{S}, \|\cdot\|_a)$ such that $f(y_1, y_2) = \frac{y_1 - y_2}{\|y_1 - y_2\|_b}$ is surjective. We show that f is Lipschitz continuous, and conclude with Lemma A.2. For $(y_1, y_2), (y_1', y_2') \in \mathcal{Q}'$, we have

$$\begin{aligned}
\|f(y_1, y_2) - f(y_1', y_2')\|_a &= \left\| \frac{y_1 - y_2}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1' - y_2'\|_b} \right\|_a \\
&\leq \left\| \frac{y_1 - y_2}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1 - y_2\|_b} \right\|_a + \left\| \frac{y_1' - y_2'}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1' - y_2'\|_b} \right\|_a.
\end{aligned}$$

Since $\|y_1 - y_2\|_b > \eta$, the first term is bounded by

$$\frac{1}{\eta}\Big(\|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a\Big) = \frac{1}{\eta}\|(y_1, y_2) - (y_1', y_2')\|_a,$$

while the second term is bounded by

$$\begin{aligned}
\|y_1' - y_2'\|_a\left|\frac{1}{\|y_1 - y_2\|_b} - \frac{1}{\|y_1' - y_2'\|_b}\right| &\overset{(102)}{\leq} B\left|\frac{\|y_1' - y_2'\|_b}{\|y_1 - y_2\|_b} - 1\right| \leq \frac{B}{\eta}|\|y_1' - y_2'\|_b - \|y_1 - y_2\|_b| \\
&\leq \frac{B}{\eta}\Big(\|y_1 - y_1'\|_b + \|y_2 - y_2'\|_b\Big), \\
&\overset{(102)}{\leq} \frac{B}{A\eta}(\|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a), \\
&= \frac{B}{A\eta}\|(y_1, y_2) - (y_1', y_2')\|_a.
\end{aligned}$$

Hence we have

$$\|f(y_1, y_2) - f(y_1', y_2')\|_a \leq \frac{1 + B/A}{\eta}\|(y_1, y_2) - (y_1', y_2')\|_a.$$

The function f is Lipschitz continuous with constant $L = (1 + B/A)/\eta$, and therefore for all $\delta > 0$:

$$\mathrm{N}(\|\cdot\|_a, \mathcal{S}, \delta) \overset{\text{Lemma A.2}}{\leq} \mathrm{N}(\|\cdot\|_a, \mathcal{Q}', \delta/L) \overset{\text{Lemma A.1}}{\leq} \mathrm{N}\left(\|\cdot\|_a, Y^2, \frac{\delta}{2L}\right) \leq \mathrm{N}^2\left(\|\cdot\|_a, Y, \frac{\delta}{4L}\right).$$

□

A.3 Mixture set

Let $(X, \|\cdot\|)$ be a vector space over $\mathbb{R}$ and $Y \subset X,\ Y \neq \emptyset$. Let $k > 0$ and $\mathcal{W} \subset \mathbb{R}^k$. For $k > 0$ and a bounded set $\mathcal{W} \subset \mathbb{R}^k,\ \mathcal{W} \neq \emptyset$, denote the mixture set

$$[Y]_{k,\mathcal{W}} = \left\{\sum_{i=1}^{k} \alpha_i y_i : \boldsymbol{\alpha} \in \mathcal{W},\ y_i \in Y\right\}. \tag{104}$$

The **radius** of a subset Y of a semi-normed vector space $(X, \|\cdot\|)$ is denoted $\|Y\| := \sup_{x \in Y}\|x\|$.

Lemma A.6. *For all $\delta > 0$ the set $[Y]_{k,\mathcal{W}}$ satisfies*

$$\mathrm{N}(\|\cdot\|, [Y]_{k,\mathcal{W}}, \delta) \leq \min_{\tau \in]0,\ 1[} \mathrm{N}\left(\|\cdot\|_1, \mathcal{W}, \frac{(1-\tau)\delta}{\|Y\|}\right) \cdot \mathrm{N}^k\left(\|\cdot\|, Y, \frac{\tau\delta}{\|\mathcal{W}\|_1}\right). \tag{105}$$

If the semi-norm $\|\cdot\|$ is indeed a norm and Y and $\mathcal{W}$ are compact, then $Y_{k,\mathcal{W}}$ is also compact.

Proof. Let $\delta > 0$ and $\tau \in]0; 1[$. Denote $\delta_1 = \tau\delta/\|\mathcal{W}\|_1$ and $\delta_2 = (1-\tau)\delta/\|Y\|$. Also denote $N_1 = \mathrm{N}(\|\cdot\|, Y, \delta_1)$ and let $\mathcal{C}_1 = \{x_1, ..., x_{N_1}\}$ be a δ_1-covering of Y. Similarly, denote $N_2 = \mathrm{N}(\|\cdot\|_1, \mathcal{W}, \delta_2)$, let $\mathcal{C}_2 = \{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_{N_2}\}$ be a δ_2-covering of $\mathcal{W}$. The cardinality of the set

$$Z = \left\{\sum_{j=1}^{k} \alpha_j x_j : x_j \in \mathcal{C}_1,\ \boldsymbol{\alpha} \in \mathcal{C}_2\right\} \tag{106}$$

is $|Z| \leq N_1^k N_2$. We will show that Z is a δ-covering of $Y_{k,\mathcal{W}}$.

Consider $y = \sum_{j=1}^k \alpha_j y_j \in Y_{k,\mathcal{W}}$. By definition, there is $\bar{\boldsymbol{\alpha}} \in \mathcal{C}_2$ so that $\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_1 \leq \delta_2$, and for all $j = 1...k$, there is $\bar{y}_j \in \mathcal{C}_1$ so that $\|y_j - \bar{y}_j\| \leq \delta_1$. Denote $\bar{y} = \sum_{j=1}^k \bar{\alpha}_j \bar{y}_j \in Z$. We have

$$\begin{aligned}\|y - \bar{y}\| =& \left\| \sum_{j=1}^k \alpha_j y_j - \sum_{j=1}^k \bar{\alpha}_j \bar{y}_j \right\| \leq \left\| \sum_{j=1}^k \alpha_j y_j - \sum_{j=1}^k \alpha_j \bar{y}_j \right\| + \left\| \sum_{j=1}^k \alpha_j \bar{y}_j - \sum_{j=1}^k \bar{\alpha}_j \bar{y}_j \right\| \\ \leq& \sum_{j=1}^k |\alpha_j| \|y_j - \bar{y}_j\| + \sum_{j=1}^k |\alpha_j - \bar{\alpha}_j| \|\bar{y}_j\| \\ \leq& \|\boldsymbol{\alpha}\|_1 \delta_1 + \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_1 \|Y\| \leq \|\mathcal{W}\|_1 \delta_1 + \delta_2 \|Y\| = \delta,\end{aligned} \tag{107}$$

and Z is indeed a δ-covering of $Y_{k,\mathcal{W}}$. Therefore, we have the bound (for all τ)

$$\mathrm{N}(\|\cdot\|, Y_{k,\mathcal{W}}, \delta) \leq |Z| \leq N_1^k N_2.$$

Furthermore, in equation (107), we have shown in particular that the embedding $(y_1, ..., y_k, \boldsymbol{\alpha}) \to \sum_{j=1}^k \alpha_j y_j$ from $Y^k \times \mathcal{W}$ to $Y_{k,\mathcal{W}}$ is continuous. Hence if Y and $\mathcal{W}$ are compact $Y_{k,\mathcal{W}}$ is the continuous image of a compact set and is compact. □

B Proofs on mixtures of distributions

We gather here all proofs related to results stated in Section 5.

B.1 Proof of Theorem 5.11

We start with the following lemma.

Lemma B.1. *Consider a kernel κ with and an integer $k \geq 1$ such that κ has its k-coherence with respect to $\mathcal{T}$ bounded by $\zeta < 1$. Then the normalized secant of the model set $\mathfrak{S}_k(\mathcal{T})$ of 2-separated mixtures is made of mixtures of $2k$ normalized dipoles:*

$$\mathcal{S}_\kappa \subset \left\{ \sum_{l=1}^{2k} \alpha_l \mu_l : \ (1+\zeta)^{-1} \leq \sum_{l=1}^{2k} \alpha_l^2 \leq (1-\zeta)^{-1}, \ \alpha_l \geq 0, \ \mu_l \in \mathcal{D}, l = 1, \ldots, 2k \right\}, \tag{108}$$

where the normalized dipoles $(\mu_l)_{1 \leq l \leq 2k}$ associated to nonzero coefficients α_l are pairwise 1-separated.

Proof. By definition any $\mu \in \mathcal{S}_\kappa$ can be written as $\mu = (\tau - \tau')/\|\tau - \tau'\|_\kappa$ with $\tau, \tau' \in \mathfrak{S}_k(\mathcal{T})$ and $\|\tau - \tau'\|_\kappa > 0$. By Lemma 5.4 we have $\tau - \tau' = \sum_{l=1}^\ell \nu_l$ where the ν_l are non-zero dipoles that are 1-separated from one another and $\ell \leq 2k$. With $\alpha_l := \frac{\|\nu_l\|_\kappa}{\left\|\sum_{i=1}^l \nu_i\right\|_\kappa} > 0$, $\mu_l := \nu_l / \|\nu_l\|_\kappa$ (note that by assumption κ is locally characteristic with respect to $\mathcal{T}$ hence $\|\nu_l\|_\kappa > 0$) we can write

$$\mu = \frac{\sum_{l=1}^\ell \nu_l}{\left\|\sum_{l=1}^\ell \nu_l\right\|_\kappa} = \sum_{l=1}^\ell \frac{\|\nu_l\|_\kappa}{\left\|\sum_{l=1}^\ell \nu_l\right\|_\kappa} \cdot \frac{\nu_l}{\|\nu_l\|_\kappa} = \sum_{l=1}^\ell \alpha_l \cdot \mu_l$$

By construction $\mu_l \in \mathcal{D}$, and by definition of $2k$-coherence we have

$$\sum_{i=1}^\ell \alpha_i^2 = \frac{\sum_{i=1}^\ell \|\nu_i\|_\kappa^2}{\left\|\sum_{i=1}^\ell \nu_i\right\|_\kappa^2} \in \left[(1+\zeta)^{-1}, (1-\zeta)^{-1}\right].$$

If needed, we iteratively add to μ arbitrary normalized dipoles μ_l (not necessarily 1-separated from another) with $\alpha_l = 0$ for $l = \ell + 1 \ldots 2k$. □

To prove Theorem 5.11 we use the following version of Bernstein's inequality.

Proposition B.2 (Massart 2007, Corollary 2.10)**.** *Let X_i, $i = 1, \dots, N$ be i.i.d. real-valued random variables. Denote $(\cdot)_+ = \max(\cdot, 0)$ and assume there exists positive numbers σ^2 and u such that*

$$\mathbb{E}(X^2) \leq \sigma^2,$$
$$\mathbb{E}((X)_+^q) \leq \frac{q!}{2}\sigma^2 u^{q-2}, \quad \text{for all integers } q \geq 3.$$

Then for any $t > 0$ we have

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} X_i \geq \mathbb{E}(X) + t\right] \leq \exp\left(\frac{-Nt^2}{2(\sigma^2 + ut)}\right).$$

For both lemmas we start from the observation that

$$\frac{\|\mathcal{A}(\tau - \tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} - 1 = \tfrac{1}{m}\sum_{j=1}^{m} Z(\omega_j) \qquad \text{with} \qquad Z(\omega) := \frac{|\langle \tau, \phi_\omega\rangle - \langle \tau', \phi_\omega\rangle|^2}{\|\tau - \tau'\|_\kappa^2} - 1$$

Proof of Theorem 5.11. We will use Proposition B.2 with $X = Y(\omega) := |\langle \tau - \tau', \phi_\omega\rangle|^2 / \|\tau - \tau'\|_\kappa^2$, hence we need to control the moments of Y.

Denoting $\mathcal{S}_\kappa$ the normalized secant set of $\mathfrak{S}_k(\mathcal{T})$, since $\mu := (\tau - \tau')/\|\tau - \tau'\|_\kappa \in \mathcal{S}_\kappa$ and κ has its $2k$-coherence bounded by $\zeta \leq 3/4$, we can apply Lemma B.1 to write μ as a mixture $\mu = \sum_{i=1}^{2k} \alpha_i \nu_i$ of normalized dipoles $\nu_i \in \mathcal{D}$ with $\|\boldsymbol{\alpha}\|_2^2 \leq (1-\zeta)^{-1} \leq 4$ and $\alpha_l \geq 0$. We have $Y(\omega) = \frac{|\langle \tau - \tau', \phi_\omega\rangle|^2}{\|\tau - \tau'\|_\kappa^2} = |\langle \mu, \phi_\omega\rangle|^2 \geq 0$ and $\mathbb{E}_{\omega\sim\Lambda} Y(\omega) = 1$. With $\beta_i := \alpha_i / \|\boldsymbol{\alpha}\|_1 \in [0, 1]$ we have $\sum_{i=1}^{2k} \beta_i = 1$. By convexity of $z \in \mathbb{C} \mapsto |z|^{2q}$ we get for $q \geq 2$:

$$Y^q(\omega) = |\langle \mu, \phi_\omega\rangle|^{2q} = \left|\sum_i \alpha_i \langle \nu_i, \phi_\omega\rangle\right|^{2q} = \|\boldsymbol{\alpha}\|_1^{2q} \cdot \left|\sum_i \beta_i \langle \nu_i, \phi_\omega\rangle\right|^{2q} \leq \|\boldsymbol{\alpha}\|_1^{2q} \cdot \sum_i \beta_i |\langle \nu_i, \phi_\omega\rangle|^{2q};$$
$$\mathbb{E}_{\omega\sim\Lambda}[Y(\omega)^q] \leq \|\boldsymbol{\alpha}\|_1^{2q} \cdot \sum_i \beta_i \mathbb{E}_{\omega\sim\Lambda}\left[|\langle \nu_i, \phi_\omega\rangle|^{2q}\right] \leq \frac{q!}{2} \cdot \|\boldsymbol{\alpha}\|_1^{2q} \lambda\gamma^{q-1}. \tag{109}$$

A direct use of Proposition B.2 with $u := \|\boldsymbol{\alpha}\|_1^2 \gamma$ and $\sigma^2 := \|\boldsymbol{\alpha}\|_1^4 \lambda\gamma$ would lead to a concentration function $c_\kappa(t) = \mathcal{O}\left(t^{-2}\|\boldsymbol{\alpha}\|_1^4 \lambda\gamma\right)$ for $t \leq 1$. Since $\|\boldsymbol{\alpha}\|_1^2 \leq 2k\|\boldsymbol{\alpha}\|_2^2 \leq 8k$ this would yield $c_\kappa(t) = \mathcal{O}\big(t^{-2}k^2\lambda\gamma\big)$ for $t \leq 1$. This is however suboptimal: since for $q = 1$ we have $\mathbb{E}_{\omega\sim\Lambda}|Y(\omega)|^q = \|\tau - \tau'\|_\kappa^2 / \|\tau - \tau'\|_\kappa^2 = 1 = (\|\boldsymbol{\alpha}\|_1^2\gamma)^{q-1}$, interpolation allows to replace $(\|\boldsymbol{\alpha}\|_1^2)^q$ in (109) by $(\|\boldsymbol{\alpha}\|_1^2)^{q-1}$ (up to log factors), as summarized by the following lemma whose proof is slightly postponed.

Lemma B.3. *Assume that the random variable $X \geq 0$ satisfies $\mathbb{E}(X) = 1$ and $\mathbb{E}[X^q] \leq \frac{q!}{2} a w^{q-1}$ for any integer $q \geq 2$, where $a \geq 2e$, $w > 0$. Then for any integer $q \geq 2$ we have*

$$\mathbb{E}[X^q] \leq \tfrac{q!}{2}\sigma^2 u^{q-2}, \tag{110}$$

with $u := w \log(ea/2)$ and $\sigma^2 := 2e \cdot w \log^2(ea/2)$.

As $\lambda \geq 1$ and $\|\boldsymbol{\alpha}\|_1^2 \leq 8k$, putting $a := 8k\lambda \geq \max(\|\boldsymbol{\alpha}\|_1^2 \lambda, 2e)$, $w := 8k\gamma \geq \gamma\|\boldsymbol{\alpha}\|_1^2$, by (109) the assumptions of Lemma B.3 are satisfied hence

$$\mathbb{E}_{\omega\sim\Lambda}[Y(\omega)^q] \leq \frac{q!}{2}\sigma^2 u^{q-2}, \tag{111}$$

with $u := w\log(ea/2)$ and $\sigma^2 := 2ew\log^2(ea/2)$. Since $a \geq 2e$, we have $\log(ea/2) \geq 2$ hence $u/\sigma^2 = (2e\log(ea/2))^{-1} \leq \frac{1}{4e} \leq \frac{1}{3}$ (we chose the factor $1/3$ for unification with the classical form of Bernstein's inequality, see e.g. Lemma 5.5 in [Gribonval et al., 2021]). Applying Proposition B.2 to $X = Y(\omega)$ and to $X = -Y(\omega)$, we get

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\tau-\tau')\|_2^2}{\|\tau-\tau'\|_\kappa^2} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{mt^2}{2\sigma^2(1+t\frac{u}{\sigma^2})}\right) \leq 2\exp\left(-\frac{mt^2}{2\sigma^2(1+t/3)}\right), \qquad \text{for each } t > 0.$$

Finally, we have $V := \sigma^2 = 16e\gamma k\log^2(4ek\lambda)$, which yields (63) . □

Proof of Lemma B.3. Consider arbitrary integers $q \geq 2$, $p \geq 2$ and a real number $1 < p'$ such that $1/p' + 1/p = 1$. By Hölder's inequality, as the integer $r := p(q-1)+1 = p(q-1+1/p) = p(q-1/p')$ satisfies $r \geq 2$, leveraging the assumptions yields

$$\begin{aligned}\mathbb{E}(X^q) = \mathbb{E}(X^{1/p'}X^{q-1/p'}) &\leq \left(\mathbb{E}(X^{1/p'})^{p'}\right)^{1/p'}\left(\mathbb{E}(X^{q-1/p'})^{p}\right)^{1/p} = (\mathbb{E}X^r)^{1/p} \\ &\leq \left(\tfrac{r!}{2}aw^{r-1}\right)^{1/p} = \left(\tfrac{r!}{2}\right)^{1/p}\cdot a^{1/p}w^{q-1}.\end{aligned}$$

As $p \geq 1$ we have $r = pq - p + 1 \leq pq$, hence

$$r! \leq (pq)! = \prod_{i=1}^{pq} i = \prod_{i=1}^{q}\prod_{j=1}^{p}(p(i-1)+j) \leq \prod_{i=1}^{q}(pi)^p = (p^q q!)^p.$$

Combining the above we obtain

$$\mathbb{E}(X^q) \leq q!\cdot p^q\cdot(a/2)^{1/p}\cdot w^{q-1}.$$

If $a > 2e$, setting $p := \lceil\log(a/2)\rceil > 1$ yields $\log(a/2) \leq p < 1 + \log(a/2) = \log(ea/2)$ and

$$p^q\cdot(a/2)^{1/p} = p^q\cdot e^{\frac{\log(a/2)}{p}} \leq (\log(ea/2))^q\cdot e.$$

We conclude that for any $q \geq 2$

$$\mathbb{E}(X^q) \leq q!\cdot(\log(ea/2))^q\cdot e\cdot w^{q-1} = \tfrac{q!}{2}\cdot[2e\cdot w\log^2(ea/2)][w\log(ea/2)]^{q-2}.$$

If $a = 2e$ we establish the same bounds with arbitrary $a' > a$ and take their infimum. □

B.2 Proof of Theorem 5.12

We prove that $\|\mathcal{D}\|_{\mathcal{F}} \geq 1$ with a minor adaptation of the arguments showing that $\|\mathcal{S}_\kappa\|_{\mathcal{F}} \geq 1$ in [Gribonval et al., 2021, Proof of Lemma 5.5]. For $\mathcal{F}$-integrable π, π' and arbitrary $\alpha, \alpha' \in \mathbb{R}$ the left hand side inequality in (64) holds since

$$\|\alpha\pi - \alpha'\pi'\|_\kappa^2 = \mathbb{E}_{\omega\sim\Lambda}|\alpha\langle\pi,\phi_\omega\rangle - \alpha'\langle\pi',\phi_\omega\rangle|^2 \leq \sup_{\omega\sim\Lambda}|\alpha\langle\pi,\phi_\omega\rangle - \alpha'\langle\pi',\phi_\omega\rangle|^2 = \|\alpha\pi - \alpha'\pi'\|_{\mathcal{F}}^2,$$

and the r.h.s. inequality in (64) holds by definition of $\|\mathcal{D}\|_{\mathcal{F}}$. Combined, they imply $\|\mathcal{D}\|_{\mathcal{F}} \geq 1$.

The rest of the proof relies on two lemmas.

Lemma B.4. *Consider a random feature family $(\{\phi_\omega\}_{\omega\in\Omega}, \Lambda)$, the induced average kernel κ, and $d_{\mathcal{F}}$ the pseudo-metric defined in (52). Let $\mathfrak{S}$ be an arbitrary model set and $\mathcal{S}_\kappa$ be its normalized secant set. For each $\delta > 0$ we have*

$$\mathrm{N}(d_{\mathcal{F}}, \mathcal{S}_\kappa, \delta) \leq \mathrm{N}\left(\|\cdot\|_{\mathcal{F}}, \mathcal{S}_\kappa, \frac{\delta}{2\|\mathcal{S}_\kappa\|_{\mathcal{F}}}\right).$$

Remark B.5. *Lemma B.4 is valid beyond the case of mixture models.*

Proof. Consider $\mu_i = (\tau_i - \tau_i')/\|\tau_i - \tau_i'\|_\kappa$, $i = 1, 2$ in $\mathcal{S}_\kappa$. By definition of $\|\mathcal{S}_\kappa\|_\mathcal{F}$, for all $\phi_\omega \in \mathcal{F}$ we have $|\langle \mu_i, \phi_\omega \rangle| \leq \|\tau_i - \tau_i'\|_\mathcal{F} / \|\tau_i - \tau_i'\|_\kappa \leq \|\mathcal{S}_\kappa\|_\mathcal{F}$, hence

$$\begin{aligned} d_\mathcal{F}(\mu_1, \mu_2) = \sup_{\omega \in \Omega} \Big| |\langle \mu_1, \phi_\omega \rangle|^2 - |\langle \mu_2, \phi_\omega \rangle|^2 \Big| &= \sup_{\omega \in \Omega} \Big| |\langle \mu_1, \phi_\omega \rangle| + |\langle \mu_2, \phi_\omega \rangle| \Big| \cdot \Big| |\langle \mu_1, \phi_\omega \rangle| - |\langle \mu_2, \phi_\omega \rangle| \Big| \\ &\leq \sup_\omega 2\|\mathcal{S}_\kappa\|_\mathcal{F} \cdot |\langle \mu_1 - \mu_2, \phi_\omega \rangle| = 2\|\mathcal{S}_\kappa\|_\mathcal{F} \|\mu_1 - \mu_2\|_\mathcal{F}. \end{aligned}$$

We conclude using Lemma A.2. □

Lemma B.6. *Consider a kernel κ on $\mathcal{Z}$, $k \geq 1$ such that κ has its $2k$-coherence bounded by $\zeta \leq 3/4$, $\mathcal{S}_\kappa$ the normalized secant set of $\mathfrak{S}_k(\mathcal{T})$, and a semi-norm $\|\cdot\|$. Then for each $\delta > 0$:*

$$\mathrm{N}(\|\cdot\|, \mathcal{S}_\kappa, \delta) \leq \Big[\mathrm{N}\Big(\|\cdot\|, \mathcal{D}, \tfrac{\delta}{8\sqrt{2k}} \Big) \cdot \max\Big(1, \tfrac{32 \|\mathcal{D}\| \cdot \sqrt{2k}}{\delta} \Big) \Big]^{2k}.$$

Proof. Denote $[Y]_{k,\mathcal{W}}$ the set of k-mixtures of elements in Y with weights in $\mathcal{W}$ (see (104)). Any $\boldsymbol{\alpha} \in \mathbb{R}^{2k}$ such that $\|\boldsymbol{\alpha}\|_2 \leq (1-\zeta)^{-1/2} \leq 2$ satisfies $\|\boldsymbol{\alpha}\|_1 \leq \sqrt{2k}\|\boldsymbol{\alpha}\|_2 \leq 2\sqrt{2k}$. As κ has $2k$-coherence bounded by $\zeta \leq 3/4$, it follows by Lemma B.1 that $\mathcal{S}_\kappa \subset [\mathcal{D}]_{2k,\mathcal{B}}$ with $\mathcal{D}$ the set of normalized dipoles and $\mathcal{B} := \mathcal{B}_{\mathbb{R}^{2k}, \|\cdot\|_1}(0, R)$ the closed ℓ^1 ball of radius $R := 2\sqrt{2k}$ in $\mathbb{R}^{2k}$.

We use generic lemmas on covering numbers that can be found in Appendix A. Exploiting Lemma A.6 will involve the following two quantities

$$\|\mathcal{B}\|_1 = R = 2\sqrt{2k}; \qquad D := \|\mathcal{D}\| = \sup_{\mu \in \mathcal{D}} \|\mu\|. \tag{112}$$

We get for each $\delta > 0$,

$$\begin{aligned} \mathrm{N}(\|\cdot\|, \mathcal{S}_\kappa, \delta) &\overset{\text{[Lemma A.1]}}{\leq} \mathrm{N}\Big(\|\cdot\|, [\mathcal{D}]_{2k,\mathcal{B}}, \tfrac{\delta}{2} \Big) \\ &\overset{\text{[Lemma A.6 with } \tau = \frac{1}{2} \text{\&(112)]}}{\leq} \mathrm{N}\big(\|\cdot\|_1, \mathcal{B}, \tfrac{\delta}{4D} \big) \cdot \mathrm{N}^{2k}\big(\|\cdot\|, \mathcal{D}, \tfrac{\delta}{4R} \big) \\ &\overset{\text{[Lemma A.4]}}{\leq} \big[\max\big(1, \tfrac{16RD}{\delta}\big) \cdot \mathrm{N}\big(\|\cdot\|, \mathcal{D}, \tfrac{\delta}{4R} \big) \big]^{2k}. \end{aligned}$$

We conclude by replacing R, D by their values from (112). □

To wrap up the proof of Theorem 5.12, we exploit Lemmas B.4 and B.6 with $\delta' = \frac{\delta}{2\|\mathcal{S}_\kappa\|_\mathcal{F}}$ to get

$$\mathrm{N}(d_\mathcal{F}, \mathcal{S}_\kappa, \delta) \leq \mathrm{N}\Big(\|\cdot\|_\mathcal{F}, \mathcal{S}_\kappa, \frac{\delta}{2\|\mathcal{S}_\kappa\|_\mathcal{F}} \Big) \leq \Big[\mathrm{N}\Big(\|\cdot\|_\mathcal{F}, \mathcal{D}, \tfrac{\delta}{16\|\mathcal{S}_\kappa\|_\mathcal{F}\sqrt{2k}} \Big) \cdot \max\Big(1, \tfrac{64\|\mathcal{D}\|_\mathcal{F} \cdot \|\mathcal{S}_\kappa\|_\mathcal{F} \sqrt{2k}}{\delta} \Big) \Big]^{2k}.$$

Combined with Theorem 5.10, this yields the result.

B.3 Proof of Theorem 5.13

By Theorem 5.10 the normalized secant $\mathcal{S}_\kappa$ of $\mathfrak{S}_k(\mathcal{T})$ satisfies

$$\|\mathcal{S}_\kappa\|_\mathcal{F} \leq \sqrt{8k}\|\mathcal{D}\|_\mathcal{F}, \qquad \|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}} \leq \sqrt{8k}\|\mathcal{D}\|_{\Delta\mathcal{L}}.$$

By [Gribonval et al., 2021, Lemma 5.5] it follows that for each $t > 0$

$$c_\kappa(t) \leq \|\mathcal{S}_\kappa\|_\mathcal{F}^2 \cdot \frac{2(1 + t/3)}{t^2} \leq 8k \cdot \|\mathcal{D}\|_\mathcal{F}^2 \cdot \frac{2(1 + t/3)}{t^2}.$$

Combining with Lemma 5.11 and using that for $t = \delta/2 \le 1/2$ we have $2(1+t/3)/t^2 \le (7/3)/(\delta/2)^2 = (28/3)/\delta^2 \le 10\delta^{-2}$ we obtain

$$c_\kappa(\delta/2) \le 10 \cdot \delta^{-2} \cdot 8k \cdot \min\Big(2e\gamma \log^2(4ek\lambda), \|\mathcal{D}\|_{\mathcal{F}}^2\Big).$$

By [Gribonval et al., 2021, Lemma 5.5] we have $\|\mathcal{S}_\kappa\|_{\mathcal{F}} \ge 1$ and by Theorem 5.12 we have $\|\mathcal{D}\|_{\mathcal{F}} \ge 1$, hence $1 \le 64k\|\mathcal{D}\|_{\mathcal{F}} \le 256k\|\mathcal{D}\|_{\mathcal{F}}^2$. For $0 < \delta < 1$, since $\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \delta) \le 2(C/\delta)^r$ we obtain

$$\max\Big(1, \tfrac{256k\|\mathcal{D}\|_{\mathcal{F}}^2}{\delta/2}\Big) = \tfrac{512k\|\mathcal{D}\|_{\mathcal{F}}^2}{\delta}$$
$$\mathrm{N}\Big(\|\cdot\|_{\mathcal{F}}, \mathcal{D}, \tfrac{\delta/2}{64k\|\mathcal{D}\|_{\mathcal{F}}}\Big) \le 2(128kC\|\mathcal{D}\|_{\mathcal{F}}/\delta)^r.$$

Taking the logarithms and using Theorem 5.12 we obtain

$$\begin{aligned}\log \mathrm{N}(d_{\mathcal{F}}, \mathcal{S}_\kappa, \delta/2) &\le 2k \cdot \Big[\log(2) + r \cdot \log(kC\|\mathcal{D}\|_{\mathcal{F}}) + r \cdot \log(128/\delta) + \log(k\|\mathcal{D}\|_{\mathcal{F}}^2) + \log(512/\delta)\Big] \\ &\le 2k(r+1)\Big[\log k + \log C + \log\|\mathcal{D}\|_{\mathcal{F}}^2 + \log(1024/\delta)\Big]\end{aligned}$$

We establish (54) and the LRIP (11) with $C_{\mathcal{A}} := \frac{8\sqrt{2k}\|\mathcal{D}\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$ and $\eta = 0$ using Theorem 5.1.

B.4 Proof of Theorem 5.15

The fact that $\|\mathcal{M}\|_{\mathcal{G}} \le \|\mathcal{D}\|_{\mathcal{G}}$ is a simple consequence of Definition (51) and the inclusion $\mathcal{M} \subset \mathcal{D}$. To prove the second inequality in (71), let $\nu = \alpha_1\pi_{\theta_1} - \alpha_2\pi_{\theta_2}$ be a nonzero dipole, which by definition means that $\theta_1 \ne \theta_2$, $\varrho(\theta_1, \theta_2) \le 1$. Consider a fixed $f \in \mathcal{G}$. We are interested in bounding $|\langle \nu, f\rangle|/\|\nu\|_\kappa$, which is invariant by rescaling ν; hence, replacing ν by $\tilde{\nu} := C_\nu^{-1}\nu$, with $C_\nu := \max(\alpha_1\|\pi_{\theta_1}\|_\kappa^{-1}, \alpha_2\|\pi_{\theta_2}\|_\kappa^{-1}) > 0$, we can assume without loss of generality that ν takes the form $\nu = s(\pi_\theta/\|\pi_\theta\|_\kappa - \alpha\pi_{\theta'}/\|\pi_{\theta'}\|_\kappa)$, with $s \in \{-1, 1\}; \alpha \in [0, 1]; \rho := \varrho(\theta, \theta') \le 1$. With this representation, since $\rho \le 1$ we get by (70)

$$\|\nu\|_\kappa^2 = \overline{\kappa}(\theta, \theta) + \alpha^2\overline{\kappa}(\theta', \theta') - 2\alpha\overline{\kappa}(\theta, \theta') \ge (1-\alpha)^2 + \alpha c\rho^2.$$

Denoting the normalized monopole $\nu_\theta := \pi_\theta/\|\pi_\theta\|_\kappa$ (and similarly θ'), we have $|\langle \nu_\theta, f\rangle| \le \|\mathcal{M}\|_{\mathcal{G}}$, and $|\langle \nu_\theta - \nu_{\theta'}, f\rangle| \le L_{\mathcal{G}}\rho$, from the assumptions of the theorem. Thus,

$$\begin{aligned}|\langle \nu, f\rangle| = |\langle \nu_\theta - \alpha\nu_{\theta'}, f\rangle| = |(1-\alpha)\langle \nu_\theta, f\rangle + \alpha\langle \nu_\theta - \nu_{\theta'}, f\rangle| &\le (1-\alpha)|\langle \nu_\theta, f\rangle| + \alpha|\langle \nu_\theta - \nu_{\theta'}, f\rangle| \\ &\le (1-\alpha)\|\mathcal{M}\|_{\mathcal{G}} + \alpha L_{\mathcal{G}}\rho.\end{aligned}$$

Gathering the two last displays, and using $c\rho^2 \le 2$, we get

$$\frac{|\langle \nu, f\rangle|}{\|\nu\|_\kappa} = \frac{(1-\alpha)\|\mathcal{M}\|_{\mathcal{G}} + \alpha L_{\mathcal{G}}\rho}{\sqrt{(1-\alpha)^2 + \alpha c\rho^2}} \le \|\mathcal{M}\|_{\mathcal{G}} + \max_{\alpha\in[0,1]} \frac{\alpha L_{\mathcal{G}}\rho}{\sqrt{(1-\alpha)^2 + \alpha c\rho^2}} \le \|\mathcal{M}\|_{\mathcal{G}} + L_{\mathcal{G}}/\sqrt{c},$$

where the last inequality follows from an elementary study of the function $\alpha \mapsto \alpha/\sqrt{(1-\alpha)^2 + a\alpha)}$ showing that it is nondecreasing on $[0, 1]$ for $a \in [0, 2]$, and therefore attains its maximum at $\alpha = 1$.

B.5 Proof of Theorem 5.16

We start with the following intermediate result:

Proposition B.7. *Consider $\mathcal{T} = (\mathbf{\Theta}, \varrho, \psi)$ a family of base distributions, κ a psd kernel on $\mathcal{Z}$ such that $\|\pi_\theta\|_\kappa > 0$ for all $\theta \in \mathbf{\Theta}$, and $\overline{\kappa}$ its associated normalized kernel on $\mathbf{\Theta}$, given by (69). Assume that κ is c-strongly characteristic, $c \in (0, 2]$. Using the shorthand $d_{ij} := \varrho(\theta_i, \theta_j)$ and $K_{ij} := \overline{\kappa}(\theta_i, \theta_j)$ for generic parameters $\theta_i, \theta_j \in \mathbf{\Theta}$, assume there is C such that the following properties hold:*

1. *if $d_{ij} \geq 1$ then $|K_{ij}| \leq C$;*
2. *if $\min(d_{ij}, d_{ik}) \geq 1$ then $|K_{ij} - K_{ik}| \leq C d_{jk}$;*
3. *if $\max(d_{ij}, d_{kl}) \leq 1$ and $\min(d_{ik}, d_{il}, d_{jk}, d_{jl}) \geq 1$ then $|K_{ik} - K_{jk} - K_{il} + K_{jl}| \leq C d_{ij} d_{kl}$.*

Then the kernel κ has mutual coherence with respect to $\mathcal{T}$ bounded by

$$M \leq \tfrac{4C}{\min(c,1)}. \tag{113}$$

Proof. Denote $\nu = \alpha_1 \frac{\pi_{\theta_1}}{\|\pi_{\theta_1}\|_\kappa} - \alpha_2 \frac{\pi_{\theta_2}}{\|\pi_{\theta_2}\|_\kappa}$ and $\nu' = \alpha_3 \frac{\pi_{\theta_3}}{\|\pi_{\theta_3}\|_\kappa} - \alpha_4 \frac{\pi_{\theta_4}}{\|\pi_{\theta_4}\|_\kappa}$ two dipoles that are 1-separated, and without loss of generality suppose that $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = a \in [0,1]$, $\alpha_4 = b \in [0,1]$. Our goal is to bound $\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa}$. Recall that $d_{ij} = \varrho(\theta_i, \theta_j)$ and $K_{ij} = \overline{\kappa}(\theta_i, \theta_j)$. We have

$$\begin{aligned}
\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} &= \frac{|K_{13} - aK_{23} - bK_{14} + abK_{24}|}{\sqrt{1 - 2aK_{12} + a^2}\sqrt{1 - 2bK_{34} + b^2}} \\
&\leq \frac{|K_{13} - K_{23} - K_{14} + K_{24}| + |(1-a)(K_{23} - K_{24})| + |(1-b)(K_{14} - K_{24})| + |(a-1)(b-1)K_{24}|}{\sqrt{(1-a)^2 + 2a(1-K_{12})}\sqrt{(1-b)^2 + 2b(1-K_{34})}}
\end{aligned}$$

By the assumptions on κ we have

$$\begin{aligned}
|K_{13} - K_{23} - K_{14} + K_{24}| &\leq C d_{12} d_{34} \quad (\text{since } d_{12} \leq 1, d_{34} \leq 1, \min(d_{13}, d_{14}, d_{23}, d_{24}) \geq 1) \\
|K_{23} - K_{24}| &\leq C d_{34} \quad (\text{since } d_{23} \geq 1 \text{ and } d_{24} \geq 1) \\
|K_{14} - K_{24}| &\leq C d_{12} \quad (\text{since } d_{14} \geq 1 \text{ and } d_{24} \geq 1) \\
|K_{24}| &\leq C \qquad (\text{since } d_{24} \geq 1) \\
2(1 - K_{12}) &\geq c d_{12}^2 \quad (\text{since } d_{12} \leq 1) \\
2(1 - K_{34}) &\geq c d_{34}^2 \quad (\text{since } d_{34} \leq 1)
\end{aligned}$$

Therefore, denoting $g(x,y) := \frac{x+y}{\sqrt{x^2 + (1-x)y^2}}$ for $0 \leq x, y \leq 1$, we have

$$\begin{aligned}
\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} &\leq C \cdot \frac{d_{12}d_{34} + (1-a)d_{34} + (1-b)d_{12} + (1-a)(1-b)}{\sqrt{(1-a)^2 + acd_{12}^2}\sqrt{(1-b)^2 + bcd_{34}^2}} \\
&= C \cdot \frac{d_{12} + 1 - a}{\sqrt{(1-a)^2 + acd_{12}^2}} \cdot \frac{d_{34} + 1 - b}{\sqrt{(1-b)^2 + bcd_{34}^2}} \\
&\leq C \cdot \frac{d_{12} + 1 - a}{\sqrt{\min(c,1)}\sqrt{(1-a)^2 + ad_{12}^2}} \cdot \frac{d_{34} + 1 - b}{\sqrt{\min(c,1)}\sqrt{(1-b)^2 + bd_{34}^2}} \\
&= \frac{C}{\min(c,1)} \cdot g(1-a, d_{12}) g(1-b, d_{34}).
\end{aligned}$$

As we have for any $0 \leq x, y \leq 1$: $g(x,y) \leq 2$ (see Lemma C.2 for a proof), gathering everything, we obtain

$$\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} \leq \frac{4C}{\min(c,1)}.$$

□

We will establish that the assumptions of Theorem 5.16 allow us to use the above proposition, for this we will also need the following technical lemmas.

Lemma B.8. *Assume that $h : \mathbb{R}_+ \to \mathbb{R}$ is differentiable and that $h'(t)$ is C-Lipschitz. Then*

$$|h(0) - h(x) - h(y) + h(x+y)| \leq xyC, \quad \forall x, y \geq 0.$$

Proof. Assume without loss of generality that $x = \min(x, y)$ and introduce $g(y) = h(y + x) - h(y)$. Notice that the considered quantity is $|g(y)-g(0)|$. By the mean value theorem, $g(y)-g(0) = g'(c)y = (h'(c + x) - h'(c))y$ for some $c \in [0, x]$, thus

$$|h(0) - h(x) - h(y) + h(x+y)| = |y(h'(c+x) - h'(c))| \leq yCx\,.$$

□

Lemma B.9. *Assume that K is differentiable with Lipschitz derivative on $[1, \infty)$ and denote $K'_{\max}, K''_{\max}$ as in the statement of Theorem 5.16. Let $(\xi_i)_{1\leq i\leq 4}$ be 4 points in a Hilbert space $\mathcal{H}$; denote $d_{ij} = \|\xi_i - \xi_j\|_{\mathcal{H}}$ and assume $d_{ij} \geq 1$ for $(i, j) \in \{(1, 3); (1, 4); (2, 3); (2, 4)\}$. Then we have*

$$|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{24})| \leq (2K'_{\max} + K''_{\max})d_{12}d_{34}. \tag{114}$$

Proof. Assume without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$ and write

$$\begin{aligned}|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{24})| \leq\ & |K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \\ & + |K(d_{24}) - K(d_{23} + d_{14} - d_{13})|. \end{aligned}\tag{115}$$

To bound the first term of the right hand side of (115), since we assumed without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$, and since $d_{13} \geq 1$ by the 1-separation assumption, we can apply Lemma B.8 with $h(t) := K(d_{13} + t)$, $x := d_{23} - d_{13} \geq 0$, $y := d_{14} - d_{13} \geq 0$, $C = K''_{\max}$, leading to

$$|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \leq K''_{\max}|(d_{23} - d_{13})(d_{14} - d_{13})| \leq 2K''_{\max}d_{12}d_{34}\,.$$

To bound the second term in (115), let $g(u) := K(\sqrt{u})$ and note that $g'(u) = K'(\sqrt{u})/2\sqrt{u}$ hence $g'(u^2) \leq K'_{\max}/2$ for $u \geq 1$. By the separation assumption we have $1 \leq d_{23} \leq d_{23} + d_{14} - d_{13}$ and $1 \leq d_{24}$. We write

$$K(d_{24}) - K(d_{23} + d_{14} - d_{13}) = g(d_{24}^2) - g((d_{23} + d_{14} - d_{13})^2) \leq \tfrac{K'_{\max}}{2}\left|d_{24}^2 - (d_{23} + d_{14} - d_{13})^2\right|,$$

where the last inequality follows from the mean value theorem. Now, it holds

$$d_{24}^2 - (d_{23} + d_{14} - d_{13})^2 = d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2 - 2(d_{13} - d_{23})(d_{13} - d_{14})\,,$$

and by the reversed triangle inequality $|d_{ij} - d_{il}| \leq d_{jl}$ for any i, j, l so that the last product is bounded in absolute value by $2d_{12}d_{34}$. It is also easy to check by expanding the squared norms $d_{ij}^2 = \|\xi_i - \xi_j\|_{\mathcal{H}}^2$ that

$$\left|d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2\right| = 2|\langle \xi_1 - \xi_2, \xi_3 - \xi_4\rangle_{\mathcal{H}}| \leq 2d_{12}d_{34}\,.$$

Gathering everything we get the desired result. □

We can now prove Theorem 5.16.

Proof of Theorem 5.16. Since $K(0) = 1$ and $K(u) \leq 1 - cu^2/2$, the kernel κ is c-strongly locally characteristic with respect to $\mathcal{T}$.

We exhibit a constant C allowing the use of Proposition B.7. Consider generic parameters θ_i, θ_j, and denote as before $d_{ij} = \varrho(\theta_i, \theta_j)$, $K_{ij} = \overline{\kappa}(\theta_i, \theta_j) = K(d_{ij})$. Since $|K(u)| \leq K_{\max}$ for $u \geq 1$ we get $|K_{ij}| \leq K_{\max}$ if $d_{ij} \geq 1$. By the mean value theorem and the reversed triangle inequality, if $\min(d_{ij}, d_{il}) \geq 1$ then as $|K'(u)| \leq K'_{\max}$ for $u \geq 1$ we get $|K_{ij} - K_{il}| = |K(d_{ij}) - K(d_{il})| \leq K'_{\max}|d_{ij} - d_{il}| \leq K'_{\max} d_{jl}$. Applying Lemma B.9 we get if $d_{12} \leq 1$, $d_{34} \leq 1$ and $\min(d_{13}, d_{14}, d_{23}, d_{24}) \geq 1$ that:

$$|K_{13} - K_{23} - K_{14} + K_{24}| \leq (2K'_{\max} + K''_{\max}) d_{12} d_{34}$$

To conclude observe that $C := \max(K_{\max}, K'_{\max}, (2K'_{\max} + K''_{\max})) = \max(K_{\max}, (2K'_{\max} + K''_{\max}))$. □

C Proofs for Section 6

We will start with an elementary result introducing a "canonical" representation of normalized dipoles.

Lemma C.1. *The set of normalized dipoles can be written as*

$$\mathcal{D} = \left\{ \frac{\nu}{\|\nu\|_\kappa} : \nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta); s \in \{-1, +1\}; 0 \leq \alpha \leq 1; 0 < \|\theta' - \theta\| \leq 1 \right\}.$$

Proof. Any element in $\mathcal{D}$ can (by definition (56)) be written as $\nu/\|\nu\|_\kappa$, where ν is a nonzero dipole of the form $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$, with $\alpha_1, \alpha_2 \geq 0$ and $\|\theta_1 - \theta_2\| \leq 1$. Let $\zeta = \max(\alpha_1, \alpha_2) > 0$ since ν is nonzero. Then $\nu' = (\zeta\|\pi_0\|_\kappa)^{-1}\nu$ is such that $\nu'/\|\nu'\|_\kappa = \nu/\|\nu\|_\kappa$, and $\nu' = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta)$ with $s \in \{-1, +1\}$; $0 \leq \alpha \leq 1$; $0 < \|\theta' - \theta\| \leq 1$. □

C.1 Proof of Lemma 6.5

From Lemma C.1, any normalized dipole can be written as $\mu = \nu/\|\nu\|_\kappa$ with $\nu = s\|\pi_0\|_\kappa^{-1}(\pi_\theta - \alpha\pi_{\theta'})$, $s \in \{-1, 1\}$, $0 \leq \alpha \leq 1$ and $x := \theta - \theta' \neq 0$, $0 < \|x\| \leq 1$. Denote $u := x/\|x\|$. Since $\|\pi_\theta\|_\kappa = \|\pi_{\theta'}\|_\kappa = \|\pi_0\|_\kappa$, reusing (78) and the definition of $\overline{\kappa}$ we have

$$\|\nu\|_\kappa^2 = \|\pi_0\|_\kappa^{-2}\Big(\|\pi_\theta\|_\kappa^2 + \alpha^2\|\pi_{\theta'}\|_\kappa^2 - 2\alpha\kappa(\pi_\theta, \pi_{\theta'})\Big) = 1 + \alpha^2 - 2\alpha\overline{\kappa}(x) = (1-\alpha)^2 + 2\alpha(1 - \overline{\kappa}(x));$$

$$\begin{aligned}\|\pi_0\|_\kappa^2 |\langle \nu, \phi_\omega\rangle|^2 &= |\langle \pi_0, \phi_\omega\rangle|^2 \cdot \left| e^{\jmath\langle\omega,\theta\rangle} - \alpha e^{\jmath\langle\omega,\theta'\rangle}\right|^2 = |\langle \pi_0, \phi_\omega\rangle|^2 \cdot \big((1-\alpha)^2 + 2\alpha(1 - \cos\langle\omega, x\rangle)\big) \\ &\leq |\langle \pi_0, \phi_\omega\rangle|^2 \cdot \max\Big(1, \tfrac{1-\cos\langle\omega,x\rangle}{1-\overline{\kappa}(x)}\Big) \cdot \big((1-\alpha)^2 + 2\alpha(1 - \overline{\kappa}(x))\big).\end{aligned}$$

Together, the last two inequalities imply

$$\|\pi_0\|_\kappa^2 |\langle \mu, \phi_\omega\rangle|^2 \leq |\langle \pi_0, \phi_\omega\rangle|^2 \cdot \max\Big(1, \tfrac{1-\cos\langle\omega,x\rangle}{1-\overline{\kappa}(x)}\Big). \tag{116}$$

By (79) we have $1 - \overline{\kappa}(x) \geq b^{-1}\min(1, (\|x\|/a)^2)$ hence:

- if $\|x\| \geq a$ then $0 \leq 1 - \cos\langle\omega, x\rangle \leq 2 = 2\min(1, (\|x\|/a)^2)) \leq 2b(1 - \overline{\kappa}(x))$ hence, since we assumed $b \geq 1/2$,
$$\max(1, \tfrac{1-\cos\langle\omega,x\rangle}{1-\overline{\kappa}(x)}) \leq \max(1, 2b) = 2b;$$

- if $\|x\| \le a$ then, since $\sin^2 t \le t^2$ for each $t \in \mathbb{R}$ we have $2\sin^2(t/2) \le \frac{t^2}{2}$ and

$$0 \le 1 - \cos\langle\omega, x\rangle = 2\sin^2\tfrac{\langle\omega,x\rangle}{2} \le \tfrac{1}{2}\langle\omega, x\rangle^2 = \tfrac{a^2}{2}\langle\omega, u\rangle^2 \cdot (\|x\|/a)^2 \le \tfrac{a^2}{2}\langle\omega, u\rangle^2 \cdot b(1-\overline{\kappa}(x)),$$

$$\text{implying} \qquad \max(1, \tfrac{1-\cos\langle\omega,x\rangle}{1-\overline{\kappa}(x)}) \le \max(1, \tfrac{ba^2}{2}\langle\omega, u\rangle^2).$$

Since $w(\omega) \ge 1$ we have $|\langle\pi_0, \phi_\omega\rangle| \le \left|\mathbb{E}_{X\sim\pi_0} e^{\jmath\langle\omega, X\rangle}\right| \le 1$, hence for any integer $q \ge 1$ we have

$$\mathbb{E}_{\omega\sim\Lambda}|\langle\pi_0, \phi_\omega\rangle|^{2q} \le \mathbb{E}_{\omega\sim\Lambda}|\langle\pi_0, \phi_\omega\rangle|^2 = \|\pi_0\|_\kappa^2. \tag{117}$$

Denoting $Y(\omega) := \|\pi_0\|_\kappa^2 |\langle\mu, \phi_\omega\rangle|^2$, we obtain by (116) and the previous cases, for any integer $q \ge 2$:

- if $\|x\| \ge a$ then $|Y(\omega)|^q \le |\langle\pi_0, \phi_\omega\rangle|^{2q} \cdot (2b)^{2q}$; by (117) we get $\mathbb{E}_{\omega\sim\Lambda}[|Y(\omega)|^q] \le \|\pi_0\|_\kappa^2 \cdot (2b)^q$;
- if $\|x\| \le a$ then $|Y(\omega)|^q \le |\langle\pi_0, \phi_\omega\rangle|^{2q} \cdot \max\Big(1, (ba^2/2)^q\langle\omega, u\rangle^{2q}\Big) \le |\langle\pi_0, \phi_\omega\rangle|^{2q} \cdot \Big(1 + (ba^2/2)^q\langle\omega, u\rangle^{2q}\Big)$; using assumption (80) and (117)we get

$$\mathbb{E}_{\omega\sim\Lambda}[|Y(\omega)|^q] \le \|\pi_0\|_\kappa^2 + (ba^2/2)^q \cdot \mathbb{E}_{\omega\sim\Lambda}\Big[|\langle\pi_0, \phi_\omega\rangle|^{2q} \cdot \langle\omega, u\rangle^{2q}\Big] \le \|\pi_0\|_\kappa^2 + (ba^2/2)^q\|\pi_0\|_\kappa^2 \frac{q!}{2}\lambda_0^q$$
$$\overset{q\ge2}{\le} \|\pi_0\|_\kappa^2 \frac{q!}{2}\big(1^q + (ba^2\lambda_0/2)^q\big) \le \|\pi_0\|_\kappa^2 \frac{q!}{2}(1 + ba^2\lambda_0/2)^q.$$

Combining both cases we obtain $\mathbb{E}_{\omega\sim\Lambda}[|Y(\omega)|^q] \le \|\pi_0\|_\kappa^2 \frac{q!}{2}\max(2b, 1 + ba^2\lambda_0/2)^q$. Finally, we get

$$\mathbb{E}_{\omega\sim\Lambda}\Big[|\langle\mu, \phi_\omega\rangle|^{2q}\Big] = (\|\pi_0\|_\kappa^{-2})^q \mathbb{E}_{\omega\sim\Lambda}[|Y(\omega)|^q] \le (\|\pi_0\|_\kappa^{-2})^{q-1} \frac{q!}{2}\max(2b, 1 + ba^2\lambda_0/2)^q.$$

C.2 Proof of Lemma 6.7

The following bound will be useful.

Lemma C.2. *For each $0 \le x, y \le 1$, $(x, y) \ne (0, 0)$ we have $g(x, y) := \frac{x+y}{\sqrt{x^2+(1-x)y^2}} \le 2$.*

Proof. Since $2xy \le x^2 + y^2$, we have

$$g^2(x, y) = \frac{(x+y)^2}{x^2 + (1-x)y^2} = 1 + \frac{2xy + xy^2}{x^2 + y^2 - xy^2} \le 1 + \frac{2xy + xy^2}{2xy - xy^2} = 1 + \frac{2+y}{2-y} = \frac{4}{2-y} \le 4.$$

□

Proof of Lemma 6.7. The argument relies on the decomposition (straightforward from the "canonical" dipole representation introduced in Lemma C.1) $\mathcal{D} = \mathcal{D}_\eta \cup \overline{\mathcal{D}}_\eta$, where (for $\eta > 0$ to be soon specified)

$$\mathcal{D}_\eta := \left\{\frac{\nu}{\|\nu\|_\kappa} : \nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta), \theta, \theta' \in \mathbf{\Theta}, \|\theta - \theta'\| \le 1, 0 \le \alpha \le 1, \|\nu\|_\kappa > \eta\right\},$$
$$\overline{\mathcal{D}}_\eta := \left\{\frac{\nu}{\|\nu\|_\kappa} : \nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta), \theta, \theta' \in \mathbf{\Theta}, \|\theta - \theta'\| \le 1, 0 \le \alpha \le 1, \|\nu\|_\kappa \le \eta\right\},$$

so that

$$\mathrm{N}(\|\cdot\|, \mathcal{D}, \delta) \le \mathrm{N}(\|\cdot\|, \mathcal{D}_\eta, \delta) + \mathrm{N}\big(\|\cdot\|, \overline{\mathcal{D}}_\eta, \delta\big). \tag{118}$$

By Theorem 5.12 we have $D := \|\mathcal{D}\|_{\mathcal{F}} \geq 1$, and we will establish below that for $\eta := \frac{\delta}{8C''_{\mathcal{F}}} > 0$:

$$\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}_\eta, \delta) \leq \max\Big(1, \tfrac{12C_{\mathcal{T}}(C_{\mathcal{F}}+C'_{\mathcal{F}}+DC''_{\mathcal{F}})}{\delta}\Big)^{4(d+1)}; \tag{119}$$

$$\mathrm{N}\big(\|\cdot\|_{\mathcal{F}}, \overline{\mathcal{D}}_\eta, \delta\big) \leq \max\Big(1, \tfrac{64C_{\mathcal{T}}(C_{\mathcal{F}}+C'_{\mathcal{F}}+C''_{\mathcal{F}})}{\delta}\Big)^{2d+1}. \tag{120}$$

It is clear that (118), (119), (120) lead to the announced estimate (83) (using $D \geq 1$).

Step 1: covering numbers of $\mathcal{D}_\eta$.

By the first part of Theorem 5.12 we can exploit Lemma A.5 with $Y := \{\alpha\pi_\theta/\|\pi_0\|_\kappa : 0 \leq \alpha \leq 1,\ \theta \in \mathbf{\Theta}\}$, $A := 1$, $B := \|\mathcal{D}\|_{\mathcal{F}} = D$, $\|\cdot\|_a = \|\cdot\|_{\mathcal{F}}$, and $\|\cdot\|_b = \|\cdot\|_\kappa$. Since $\mathcal{D}_\eta := \Big\{\frac{y-y'}{\|y-y'\|_\kappa}, (y,y') \in \mathcal{Q}, \|y-y'\|_\kappa > \eta\Big\}$ with $\mathcal{Q} := \mathcal{Q}_1 \cup \mathcal{Q}_2$ and

$$\mathcal{Q}_1 := \{(\pi_{\theta'}, \alpha\pi_\theta)/\|\pi_0\|_\kappa, \theta, \theta' \in \mathbf{\Theta}, \|\theta-\theta'\| \leq 1, 0 \leq \alpha \leq 1\},$$
$$\mathcal{Q}_2 := \{(\alpha\pi_{\theta'}, \pi_\theta)/\|\pi_0\|_\kappa, \theta, \theta' \in \mathbf{\Theta}, \|\theta-\theta'\| \leq 1, 0 \leq \alpha \leq 1\},$$

we obtain

$$\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}_\eta, \delta) \leq \mathrm{N}^2\Big(\|\cdot\|_{\mathcal{F}}, Y, \tfrac{\delta\eta}{4(1+B/A)}\Big) \overset{B/A=D\geq 1}{\leq} \mathrm{N}^2\Big(\|\cdot\|_{\mathcal{F}}, Y, \tfrac{\delta\eta}{8D}\Big) = \mathrm{N}^2\Big(\|\cdot\|_{\mathcal{F}}, Y, \tfrac{\delta^2}{64DC''_{\mathcal{F}}}\Big). \tag{121}$$

Denoting $\mathcal{W} = [0,\ 1]$, we have $Y = [\psi(\mathbf{\Theta})]_{1,\mathcal{W}}$ (using the notation of (104)), where $\psi : \theta \mapsto \pi_\theta/\|\pi_\theta\|_\kappa = \pi_\theta/\|\pi_0\|_\kappa$. As $\|\mathcal{W}\|_1 = 1$ and $\|\psi(\mathbf{\Theta})\|_{\mathcal{F}} \leq C_{\mathcal{F}}$, by Lemma A.6 with $\tau = 1/2$ we get

$$\mathrm{N}\Big(\|\cdot\|_{\mathcal{F}}, Y, \tfrac{\delta^2}{64DC''_{\mathcal{F}}}\Big) \leq \mathrm{N}\Big(\|\cdot\|_1, \mathcal{W}, \tfrac{\delta^2}{128DC''_{\mathcal{F}}C_{\mathcal{F}}}\Big) \cdot \mathrm{N}\Big(\|\cdot\|_{\mathcal{F}}, \psi(\mathbf{\Theta}), \tfrac{\delta^2}{128DC''_{\mathcal{F}}}\Big). \tag{122}$$

As $\mathcal{W} = \mathcal{B}_{\mathbb{R}^1,|\cdot|_1}(1/2, 1/2)$, by Lemma A.4 we get $\mathrm{N}\Big(\|\cdot\|_1, \mathcal{W}, \frac{\delta^2}{128DC''_{\mathcal{F}}C_{\mathcal{F}}}\Big) \leq \max\Big(1, \frac{256DC''_{\mathcal{F}}C_{\mathcal{F}}}{\delta^2}\Big)$. Moreover, from Lemma 6.4, ψ is $L_{\mathcal{F}} = C'_{\mathcal{F}}$-Lipschitz with respect to $\|\cdot\|$ and $\|\cdot\|_{\mathcal{F}}$, thus, by Lemma A.2 and assumption (82):

$$\mathrm{N}\Big(\|\cdot\|_{\mathcal{F}}, \psi(\mathbf{\Theta}), \tfrac{\delta^2}{128DC''_{\mathcal{F}}}\Big) \leq \mathrm{N}\Big(\|\cdot\|, \mathbf{\Theta}, \tfrac{\delta^2}{128DC''_{\mathcal{F}}C'_{\mathcal{F}}}\Big) \leq \max\Big(1, \tfrac{128C_{\mathcal{T}}DC''_{\mathcal{F}}C'_{\mathcal{F}}}{\delta^2}\Big)^d$$

Combining the above we obtain (using $D \geq 1, C_{\mathcal{T}} \geq 1$; and $2ab \leq (a+b)^2$ with $a = DC''_{\mathcal{F}}, b = C'_{\mathcal{F}}+C_{\mathcal{F}}$)

$$\begin{aligned}
\mathrm{N}(\|\cdot\|_{\mathcal{F}}, \mathcal{D}_\eta, \delta) &\leq \Big[\max\Big(1, \tfrac{256DC''_{\mathcal{F}}C_{\mathcal{F}}}{\delta^2}\Big)\Big]^2 \cdot \max\Big(1, \tfrac{128C_{\mathcal{T}}DC''_{\mathcal{F}}C'_{\mathcal{F}}}{\delta^2}\Big)^{2d} \\
&\leq \max\Big(1, \tfrac{256C_{\mathcal{T}}DC''_{\mathcal{F}}(C'_{\mathcal{F}}+C_{\mathcal{F}})}{\delta^2}\Big)^{2(d+1)} \\
&\leq \max\Big(1, \tfrac{12C_{\mathcal{T}}(DC''_{\mathcal{F}}+C'_{\mathcal{F}}+C_{\mathcal{F}})}{\delta}\Big)^{4(d+1)},
\end{aligned}$$

i.e. we have obtained (119).

Step 2: local tangent approximation of $\overline{\mathcal{D}}_\eta$. To control $\mathrm{N}\big(\|\cdot\|_{\mathcal{F}}, \overline{\mathcal{D}}_\eta, \delta\big)$, the principle will be to approximate $\overline{\mathcal{D}}_\eta$ by an appropriate "tangent space", then use Lemma A.3.

To this end, let E denote the algebraic dual of smooth functions that are bounded with bounded derivatives on $\mathbb{R}^d$. The semi-norm $\|\cdot\|_{\mathcal{F}}$ is extended naturally to E as $\|\mu\|_{\mathcal{F}} := \sup_\omega |\langle\mu, \phi_\omega\rangle| \in [0, \infty]$ for any $\mu \in E$; let $\widetilde{E} := \{\mu \in E : \|\mu\|_{\mathcal{F}} < \infty\}$. Note that all finite signed measures are elements of $\widetilde{E}$. Given $\theta, \Delta \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$, define $\xi = \xi_{\theta,\Delta,\beta} \in \widetilde{E}$ by its action on functions $g : \mathbb{R}^d \to \mathbb{C}$ that are bounded with bounded gradient:

$$\langle\xi, g\rangle := \|\pi_0\|_\kappa^{-1} \cdot \mathbb{E}_{X\sim\pi_\theta}\{\langle\nabla g(X), \Delta\rangle + \beta g(X)\}.$$

Let $\mathcal{B}$ be the ball of radius 2 in $\mathbb{R}^d \times \mathbb{R}$ equipped with the norm $\|(\Delta,\beta)\|_{\texttt{mix}} := \|\Delta\| + |\beta|$.

Consider $\nu = \|\pi_0\|_\kappa^{-1} s(\pi_{\theta'} - \alpha\pi_\theta)$ with $s \in \{-1,+1\}$, $0 \le \alpha \le 1$ and $0 < \|\theta' - \theta\| \le 1$, and denote $t := \|\nu\|_\kappa$. We will show that there exists $(\Delta,\beta) \in \mathcal{B}$ such that $\mu := \nu/\|\nu\|_\kappa$ satisfies

$$\|\mu - \xi_{\theta,\Delta,\beta}\|_{\mathcal{F}} \le C''_{\mathcal{F}} \|\theta' - \theta\| \le 2C''_{\mathcal{F}} t. \tag{123}$$

Using this approximation, and the fact that to approximate any element of $\overline{\mathcal{D}}_\eta$ we can assume $t \le \eta = \delta/(8C''_{\mathcal{F}})$, we apply Lemma A.3 (with $Z = \overline{\mathcal{D}}_\eta, Y = \{\xi_{\theta,\Delta,\beta}, \theta \in \mathbf{\Theta}, (\Delta,\beta) \in \mathcal{B}\}$, $\delta' = \varepsilon = \delta/4$) to obtain

$$\mathrm{N}\big(\|\cdot\|_{\mathcal{F}}, \overline{\mathcal{D}}_\eta, \delta\big) = \mathrm{N}\big(\|\cdot\|_{\mathcal{F}}, \overline{\mathcal{D}}_\eta, 2(\delta'+\varepsilon)\big) \le \mathrm{N}\big(\|\cdot\|_{\mathcal{F}}, \{\xi_{\theta,\Delta,\beta}, \theta \in \mathbf{\Theta}, (\Delta,\beta) \in \mathcal{B}\}, \tfrac{\delta}{4}\big). \tag{124}$$

We now prove (123). Since κ is shift-invariant and locally characteristic on $\mathcal{T}$ by Proposition 6.2 we have $\|\pi_\theta\|_\kappa = \|\pi_{\theta'}\|_\kappa = \|\pi_0\|_\kappa > 0$. Denote $x := \theta' - \theta$. Since κ is 1-strongly locally characteristic we have

$$t^2 = \|\pi_0\|_\kappa^{-2} \|\pi_{\theta'} - \alpha\pi_\theta\|_\kappa^2 = 1 + \alpha^2 - 2\alpha\overline{\kappa}(\theta',\theta) = (1-\alpha)^2 + 2\alpha(1 - \overline{\kappa}(\theta',\theta)) \ge (1-\alpha)^2 + \alpha\|x\|^2.$$

Setting $\beta := s(1-\alpha)/t$ and $\Delta := sx/t$ we get using Lemma C.2

$$\|\Delta\| + |\beta| = \frac{1 - \alpha + \|x\|}{t} \le \frac{(1-\alpha) + \|x\|}{\sqrt{(1-\alpha)^2 + \alpha\|x\|^2}} = g(1-\alpha, \|x\|) \le 2.$$

Since $|\langle \omega, \Delta\rangle| \le \|\omega\|_\star \|\Delta\|$ and $\|x\|^2/t = (\|x\|/t)\|x\| = \|\Delta\|\|x\| \le 2\|x\|$, by a Taylor expansion with integral remainder term we obtain

$$\begin{aligned}
\Big|s(e^{\jmath\langle\omega,x\rangle} - 1)/t - \jmath\langle\omega,\Delta\rangle\Big| &= \Big|(e^{\jmath\langle\omega,x\rangle} - 1)/t - \jmath\langle\omega, s\Delta\rangle\Big| = \Big|(e^{\jmath t\langle\omega,s\Delta\rangle} - 1)/t - \jmath\langle\omega, s\Delta\rangle\Big| \\
&\le \sup_{0\le\tau\le t} \left|\frac{d^2}{dt^2} e^{\jmath\tau\langle\omega,s\Delta\rangle}\right| \cdot \frac{t}{2} = \langle\omega, s\Delta\rangle^2 \frac{t}{2} \le \|\omega\|_\star^2 \|\Delta\|^2 \frac{t}{2} = \|\omega\|_\star^2 \frac{\|x\|^2}{2t} \\
&\le \|\omega\|_\star^2 \|x\|.
\end{aligned} \tag{125}$$

For each ω, since $\langle \pi_\theta, \phi_\omega\rangle = e^{\jmath\langle\omega,\theta\rangle}\langle\pi_0,\phi_\omega\rangle$ we have

$$\|\pi_0\|_\kappa \langle \nu, \phi_\omega\rangle e^{-\jmath\langle\omega,\theta\rangle} = s\langle \pi_x - \alpha\pi_0, \phi_\omega\rangle = s\langle\pi_0,\phi_\omega\rangle\Big(e^{\jmath\langle\omega,x\rangle} - 1 + 1 - \alpha\Big) = t\langle\pi_0,\phi_\omega\rangle\Big(s(e^{\jmath\langle\omega,x\rangle} - 1)/t + \beta\Big).$$

Since ϕ_ω and its gradient $\nabla\phi_\omega = \phi_\omega \cdot \jmath\omega$ are bounded on $\mathbb{R}^d$, with $\xi := \xi_{\theta,\Delta,\beta} \in E$ we have

$$\begin{aligned}
\|\pi_0\|_\kappa \|\nu\|_\kappa \langle \xi, \phi_\omega\rangle e^{-\jmath\langle\omega,\theta\rangle} &= t\|\pi_0\|_\kappa \langle \xi, \phi_\omega\rangle e^{-\jmath\langle\omega,\theta\rangle} = t \cdot \mathbb{E}_{X\sim\pi_\theta}\{(\jmath\langle\omega,\Delta\rangle + \beta)\phi_\omega(X)\} e^{-\jmath\langle\omega,\theta\rangle} \\
&= t\langle\pi_0,\phi_\omega\rangle \cdot (\jmath\langle\omega,\Delta\rangle + \beta),
\end{aligned}$$

thus from the last two displays and (125) we obtain

$$\|\pi_0\|_\kappa |\langle \nu - \|\nu\|_\kappa \xi, \phi_\omega\rangle| = t|\langle\pi_0,\phi_\omega\rangle| \cdot \Big|s(e^{\jmath\langle\omega,x\rangle} - 1)/t - \jmath\langle\omega,\Delta\rangle\Big| \le t|\langle\pi_0,\phi_\omega\rangle| \|\omega\|_\star^2 \|x\|.$$

Dividing both hand sides by $\|\pi_0\|_\kappa \|\nu\|_\kappa = t\|\pi_0\|_\kappa$ and taking the supremum over ω yields

$$\|\mu - \xi\|_{\mathcal{F}} \le \|\pi_0\|_\kappa^{-1} \sup_\omega \Big(|\langle\pi_0,\phi_\omega\rangle| \|\omega\|_\star^2\Big)\|x\| = \|\pi_0\|_\kappa^{-1} \|\pi_0\|_{\mathcal{F}''} \|x\| = C''_{\mathcal{F}} \|x\|.$$

We conclude using that $\|\theta' - \theta\| = \|x\| = \|\Delta\| t \le 2t$.

Step 3: δ-covering of $\{\xi_{\theta,\Delta,\beta}\}$. Define $\delta_1 := \delta/(4(C''_{\mathcal{F}}+C'_{\mathcal{F}}))$, $\delta_2 := \delta/(2(C'_{\mathcal{F}}+C_{\mathcal{F}}))$, and consider $\mathcal{C}_1$ a δ_1-cover of $\mathbf{\Theta}$ with respect to $\|\cdot\|$ and $\mathcal{C}_2$ a δ_2-cover of $\mathcal{B}$ with respect to $\|\cdot\|_{\mathtt{mix}}$. In order to exploit (124), we now show that $\Big\{\xi_{\tilde\theta,\Delta',\beta'}, \tilde\theta \in \mathcal{C}_1, (\Delta',\beta') \in \mathcal{C}_2\Big\}$ is a δ-covering of $\{\xi_{\theta,\Delta,\beta}, \theta \in \mathbf{\Theta}, (\Delta,\beta) \in \mathcal{B}\}$.

Given any $\theta \in \mathbf{\Theta}$, $(\Delta,\beta) \in \mathcal{B}$ there are $\tilde\theta \in \mathcal{C}_1$, $(\Delta',\beta') \in \mathcal{C}_2$ such that $\Big\|\tilde\theta - \theta\Big\| \leq \delta_1$, and $\|(\Delta',\beta') - (\Delta,\beta)\|_{\mathtt{mix}} \leq \delta_2$. Observe that $\Big|e^{\jmath\langle\omega,\tilde\theta-\theta\rangle} - 1\Big| \leq \Big|\Big\langle\omega,\tilde\theta-\theta\Big\rangle\Big| \leq \|\omega\|_\star\Big\|\tilde\theta-\theta\Big\|$, and

$$|\langle \jmath\omega, \Delta\rangle + \beta| \leq \|\omega\|_\star\|\Delta\| + |\beta| \leq \max(\|\omega\|_\star, 1)\cdot\|(\Delta,\beta)\|_{\mathtt{mix}} \leq 2\max(\|\omega\|_\star,1) \leq 2(\|\omega\|_\star+1).$$

Since $\|\pi_0\|_\kappa\langle\xi_{\theta,\Delta,\beta},\phi_\omega\rangle = \langle\pi_0,\phi_\omega\rangle e^{\jmath\langle\omega,\theta\rangle}(\jmath\langle\omega,\Delta\rangle+\beta)$ (and similarly with $\tilde\theta$, Δ', β') we get

$$\begin{aligned}\|\pi_0\|_\kappa\Big|\Big\langle\xi_{\theta,\Delta,\beta} - \xi_{\tilde\theta,\Delta,\beta},\phi_\omega\Big\rangle\Big| &= |\langle\pi_0,\phi_\omega\rangle|\cdot\Big|\Big(e^{\jmath\langle\omega,\tilde\theta-\theta\rangle}-1\Big)(\jmath\langle\omega,\Delta\rangle+\beta)\Big| \\ &\leq |\langle\pi_0,\phi_\omega\rangle|\cdot\|\omega\|_\star\cdot\Big\|\tilde\theta-\theta\Big\|\cdot 2(\|\omega\|_\star+1) \\ &\leq |\langle\pi_0,\phi_\omega\rangle|\cdot(\|\omega\|_\star^2+\|\omega\|_\star)\cdot 2\delta_1,\end{aligned}$$

so that

$$\|\pi_0\|_\kappa\Big\|\xi_{\theta,\Delta,\beta}-\xi_{\tilde\theta,\Delta,\beta}\Big\|_{\mathcal{F}} \leq \sup_\omega\Big\{|\langle\pi_0,\phi_\omega\rangle|\cdot(\|\omega\|_\star^2+\|\omega\|_\star)\Big\}\cdot 2\delta_1 \leq (\|\pi_0\|_{\mathcal{F}''}+\|\pi_0\|_{\mathcal{F}'})2\delta_1. \tag{126}$$

On the other hand,

$$\begin{aligned}\|\pi_0\|_\kappa\Big|\Big\langle\xi_{\tilde\theta,\Delta,\beta} - \xi_{\tilde\theta,\Delta',\beta'},\phi_\omega\Big\rangle\Big| &= |\langle\pi_0,\phi_\omega\rangle|\cdot|\jmath\langle\omega,(\Delta'-\Delta)\rangle + (\beta'-\beta)| \\ &\leq |\langle\pi_0,\phi_\omega\rangle|\cdot(\|\omega\|_\star+1)\cdot\|(\Delta',\beta')-(\Delta,\beta)\|_{\mathtt{mix}} \\ &\leq |\langle\pi_0,\phi_\omega\rangle|\cdot(\|\omega\|_\star+1)\cdot\delta_2,\end{aligned}$$

so that

$$\|\pi_0\|_\kappa\Big\|\xi_{\tilde\theta,\Delta,\beta}-\xi_{\tilde\theta,\Delta',\beta'}\Big\|_{\mathcal{F}} \leq \sup_\omega\{|\langle\pi_0,\phi_\omega\rangle|\cdot(\|\omega\|_\star+1)\}\cdot\delta_2 \leq (\|\pi_0\|_{\mathcal{F}'}+1)\delta_2. \tag{127}$$

By a triangle inequality we combine (126)-(127) to get

$$\Big\|\xi_{\theta,\Delta,\beta}-\xi_{\tilde\theta,\Delta',\beta'}\Big\|_{\mathcal{F}} \leq (C''_{\mathcal{F}}+C'_{\mathcal{F}})2\delta_1 + (C'_{\mathcal{F}}+C_{\mathcal{F}})\delta_2 = \delta.$$

To conclude this step, we have established that

$$\begin{aligned}\mathrm{N}(\|\cdot\|_{\mathcal{F}},\{\xi_{\theta,\Delta,\beta},\theta\in\mathbf{\Theta},(\Delta,\beta)\in\mathcal{B}\},\delta) &\leq \mathrm{N}\Big(\|\cdot\|,\mathbf{\Theta},\tfrac{\delta}{4(C''_{\mathcal{F}}+C'_{\mathcal{F}})}\Big)\mathrm{N}\Big(\|\cdot\|_{\mathtt{mix}},\mathcal{B},\tfrac{\delta}{2(C'_{\mathcal{F}}+C_{\mathcal{F}})}\Big) \\ &\leq \max\Big(1,\tfrac{4C_{\mathcal{T}}(C''_{\mathcal{F}}+C'_{\mathcal{F}})}{\delta}\Big)^d\max\Big(1,\tfrac{16(C'_{\mathcal{F}}+C_{\mathcal{F}})}{\delta}\Big)^{d+1} \\ &\leq \max\Big(1,\tfrac{16C_{\mathcal{T}}(C_{\mathcal{F}}+C'_{\mathcal{F}}+C''_{\mathcal{F}})}{\delta}\Big)^{2d+1},\end{aligned} \tag{128}$$

using assumption (82) and Lemma A.4 for the second estimate, since $\mathcal{B}$ is a ball of radius 2 with respect to $\|\cdot\|_{\mathtt{mix}}$ in $\mathbb{R}^{d+1}$; and finally $C_{\mathcal{T}} \geq 1$ for the last estimate. Plugging in (128) into (124) yields (120), and the proof is done. □

C.3 Kernel mean embedding for Gaussians

The following lemma characterizes the mean map kernel on any pair of Gaussians.

Lemma C.3. *Consider a Gaussian kernel $\kappa_{\mathbf{R}}(x,x') := \exp\Big(-\frac{1}{2}\|x-x'\|^2_{\mathbf{R}}\Big)$, where $\mathbf{R}$ is an arbitrary invertible covariance matrix. For any two Gaussians $\pi_1 = \mathcal{N}(\theta_1, \mathbf{\Sigma}_1)$, $\pi_2 = \mathcal{N}(\theta_2, \mathbf{\Sigma}_2)$, the mean kernel defined from $\kappa_{\mathbf{R}}$ using* (47) *is*

$$\kappa_{\mathbf{R}}(\pi_1,\pi_2) = \frac{\sqrt{\det(\mathbf{R})}}{\sqrt{\det(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2+\mathbf{R})}} \exp\Big(-\tfrac{1}{2}\|\theta_1-\theta_2\|^2_{\mathbf{\Sigma}_1+\mathbf{\Sigma}_2+\mathbf{R}}\Big). \tag{129}$$

Proof. As $\kappa_{\mathbf{R}}(x,x') = \sqrt{\det(2\pi\mathbf{R})}\cdot \pi_{\mathbf{R}}(x-x')$ where $\pi_{\mathbf{R}} = \mathcal{N}(0,\mathbf{R})$, we have

$$\kappa_{\mathbf{R}}(\pi_1,\pi_2) = \sqrt{\det(2\pi\mathbf{R})}\int_x \pi_1(x) \underbrace{\Big(\int_{x'} \pi_2(x')\pi_{\mathbf{R}}(x-x')dx'\Big)}_{\pi_2\star\pi_{\mathbf{R}} = \mathcal{N}(\theta_2,\mathbf{\Sigma}_2+\mathbf{R}) =: \mathcal{N}(\theta_3,\mathbf{\Sigma}_3)=\pi_3} dx$$

We conclude using a property on products of Gaussians [Ahrendt, 2005, Equation (5.6)].

$$\int \pi_1(x)\pi_3(x)dx = \frac{1}{\sqrt{\det(2\pi(\mathbf{\Sigma}_1+\mathbf{\Sigma}_3))}} \exp\Big(-\tfrac{1}{2}\|\theta_1-\theta_3\|^2_{\mathbf{\Sigma}_1+\mathbf{\Sigma}_3}\Big). \qquad \square$$

C.4 Proof of Lemma 6.10

Denote $K = K_\sigma$ for brevity. If $u \geq \sigma$ then $1-K(u) \geq 1-e^{-1/2} \approx 0.39 > 1/3$. Now, if $0 < u \leq \sigma$: by concavity of the function $t \mapsto 1-e^{-t/2\sigma^2}$ on the interval $[0,\sigma^2]$, we have

$$1-e^{-t/2\sigma^2} \geq (t/\sigma^2)\cdot(1-e^{-1/2})/ > t/3\sigma^2$$

for $0 \leq t \leq \sigma^2$, hence with $t = u^2$ we get $1-K(u) \geq u^2/3\sigma^2$. This shows that $1-K(u) \geq \min(1,(u/\sigma)^2)/3$.

If $\sigma^2 \leq 1/2$ then $t \mapsto h(t) := (1-t/2)\exp(\frac{t}{2\sigma^2})$ is non-decreasing on $[0,1]$ with $h(0)=1$. For $0 \leq u \leq 1$ we obtain $(1-u^2/2)/K(u) = h(u^2) \geq 1$ hence $K(u) \leq 1-u^2/2$.

We have $K'(u) = -\frac{u}{\sigma^2}\exp(-\frac{u^2}{2\sigma^2})$, $K''(u) = (\frac{u^2}{\sigma^2}-1)\exp(-\frac{u^2}{2\sigma^2})/\sigma^2$, $K'''(u) = (3-\frac{u^2}{\sigma^2})\frac{u}{\sigma^4}\exp(-\frac{u^2}{2\sigma^2})$. Since $\sigma^2 \leq 1/4$, for $u \geq 1 \geq \sigma$ we have $K''(u) \geq 0$. Hence, K' is negative and increasing on $[1,\infty)$ and we get $K'_{\max} = |K'(1)| = \exp(-\frac{1}{2\sigma^2})/\sigma^2$. Since $K'_{\max} > K_{\max} = K(1)$ we have (cf (74))

$$C(K) = \max(K_{\max}, 2K'_{\max}+K''_{\max})) \leq 2(K'_{\max}+K''_{\max}).$$

Similarly, since $\sigma^2 \leq 1/3$, for $u \geq 1 \geq \sqrt{3}\sigma$ we have $K'''(u) \leq 0$ hence K'' is positive decreasing on $[1,\infty)$ and $K''_{\max} = K''(1) = \frac{1}{\sigma^2}\big(\frac{1}{\sigma^2}-1\big)\exp(-\frac{1}{2\sigma^2})$. As a result $K'_{\max}+K''_{\max} = \frac{1}{\sigma^4}\exp(-\frac{1}{2\sigma^2})$ and $C(K) \leq \frac{2}{\sigma^4}\exp(-\frac{1}{2\sigma^2})$.

Given $c \geq 2$, putting $\sigma_k^\star := (\sqrt{2c\log(ek)})^{-1} \in [0,\frac{1}{2}]$, since the function $t \mapsto t^2\exp(-t/2)$ is nonincreasing for $t \geq 4$, it holds for any $\sigma \leq \sigma_k^\star$ that

$$\begin{aligned} C(K) \leq 2g(1/\sigma^2) \leq 2g(1/(\sigma_k^\star)^2) &= 2(\sigma_k^\star)^{-4}\exp\Big(-\tfrac{1}{2(\sigma_k^\star)^2}\Big) = 8c^2\cdot\log^2(ek)\cdot(ek)^{-c} \\ &= \frac{8c^2}{2k-1}\cdot\log^2(ek)\cdot(ek)^{-c}(2k-1) \leq \frac{8c^2e^{-c}}{2k-1}, \end{aligned}$$

where we used at the last inequality that the function $k \mapsto \log^2(ek)\cdot(ek)^{-c}(2k-1)$ is nonincreasing for $k \geq 1$ if $c \geq 4$. The choice $c = 8$ leads to $C(K) \leq \frac{3}{16(2k-1)}$.

C.5 Proof of Equation (96)

For Diracs since $\|\cdot\| = \|\cdot\|_2/\varepsilon$ and $\|u\| = 1$ we write $u = \varepsilon v$ where $\|v\|_2 = 1$. With the probability distribution Λ on ω from (84), since $|\langle \pi_0, \phi_\omega\rangle| = 1/w(\omega) \leq 1$, $\mathbf{\Gamma} = s^{-2}\mathbf{I}_d$ (see Definition 6.9) and $C_\Lambda^{-2} = \|\pi_0\|_\kappa^2$ (see (87)), we obtain

$$\begin{array}{rcl}
\mathbb{E}_{\omega\sim\Lambda}\Big\{|\langle \pi_0, \phi_\omega\rangle|^{2q}\langle \omega, u\rangle^{2q}\Big\} & = & \displaystyle\int_{\mathbb{R}^d} w^{-2q}(\omega)\langle \omega, \varepsilon v\rangle^{2q} C_\Lambda^{-2} w^2(\omega) p_{\mathcal{N}(0,s^{-2}\mathbf{I}_d)}(\omega)d\omega \\
& \overset{w\geq 1, q\geq 1}{\leq} & \|\pi_0\|_\kappa^2 \varepsilon^{2q}\displaystyle\int_{\mathbb{R}^d}\langle \omega, v\rangle^{2q} p_{\mathcal{N}(0,s^{-2}\mathbf{I}_d)}(\omega)d\omega \\
& = & \|\pi_0\|_\kappa^2 \varepsilon^{2q} s^{-2q}\cdot \mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)}\langle s\omega, v\rangle^{2q} \\
& \overset{\omega' := s\omega}{=} & \|\pi_0\|_\kappa^2 (\varepsilon/s)^{2q}\cdot \mathbb{E}_{\omega'\sim\mathcal{N}(0,\mathbf{I}_d)}\langle \omega', u\rangle^{2q} \overset{(*)}{=} \|\pi_0\|_\kappa^2 (\varepsilon/s)^{2q}\cdot \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\xi^{2q}
\end{array}$$

where in (*) we use that as $\|v\|_2 = 1$ and $\omega' \sim \mathcal{N}(0, \mathbf{I}_d)$, $\xi := \langle \omega', u\rangle$ is standard Gaussian.

For Gaussians since $\|\cdot\| = \|\cdot\|_{\mathbf{\Sigma}}/\varepsilon$ and $\|u\| = 1$ we write $u = \varepsilon\mathbf{\Sigma}^{1/2}v$ where $\|v\|_2 = 1$. For $q \geq 1$ we have $(1 + 2qs^{-2})^{-d/2} \leq (1 + 2s^{-2})^{-d/2} = \|\pi_0\|_\kappa^2$ (see (87)). Since $|\langle \pi_0, \phi_\omega\rangle| := e^{-\omega^T\mathbf{\Sigma}\omega/2} \leq 1$ and $\Lambda(\omega) := p_{\mathcal{N}(0,s^{-2}\mathbf{\Sigma}^{-1})}(\omega)$ (see (84) and Definition 6.9), we obtain

$$\begin{array}{rcl}
|\langle \pi_0, \phi_\omega\rangle|^{2q}\Lambda(\omega) & = & \sqrt{\det(2\pi s^2\mathbf{\Sigma})}e^{-(2q+s^2)\omega^T\mathbf{\Sigma}\omega/2} = \frac{\sqrt{\det(2\pi s^2\mathbf{\Sigma})}}{\sqrt{\det(2\pi(2q+s^2)\mathbf{\Sigma})}}p_{\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{\Sigma}^{-1})}(\omega) \\
& = & (1 + 2qs^{-2})^{-d/2}p_{\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{\Sigma}^{-1})}(\omega) \leq \|\pi_0\|_\kappa^2\cdot p_{\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{\Sigma}^{-1})}(\omega),
\end{array}$$

hence

$$\begin{array}{rcl}
\mathbb{E}_{\omega\sim\Lambda}|\langle \pi_0, \phi_\omega\rangle|^{2q}\langle \omega, u\rangle^{2q} & := & \displaystyle\int_{\mathbb{R}^d}|\langle \pi_0, \phi_\omega\rangle|^{2q}\Big\langle \omega, \varepsilon\mathbf{\Sigma}^{1/2}v\Big\rangle^{2q}\Lambda(\omega)d\omega \\
& \leq & \|\pi_0\|_\kappa^2\varepsilon^{2q}\displaystyle\int_{\mathbb{R}^d}\Big\langle \mathbf{\Sigma}^{1/2}\omega, v\Big\rangle^{2q} p_{\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{\Sigma}^{-1})}(\omega)d\omega \\
& = & \|\pi_0\|_\kappa^2\varepsilon^{2q}\cdot\mathbb{E}_{\omega\sim\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{\Sigma}^{-1})}\Big\langle \mathbf{\Sigma}^{1/2}\omega, v\Big\rangle^{2q} \\
& \overset{\omega' := \mathbf{\Sigma}^{1/2}\omega}{=} & \|\pi_0\|_\kappa^2\varepsilon^{2q}\cdot\mathbb{E}_{\omega'\sim\mathcal{N}(0,(2q+s^2)^{-1}\mathbf{I}_d)}\langle \omega', v\rangle^{2q} \\
& \overset{\omega=\sqrt{2q+s^2}\omega'}{=} & \|\pi_0\|_\kappa^2\varepsilon^{2q}(\sqrt{2q+s^2})^{-2q}\cdot\mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I}_d)}\langle \omega, v\rangle^{2q} \\
& \overset{q\geq 1,(*)}{\leq} & \|\pi_0\|_\kappa^2(\varepsilon/\sqrt{2+s^2})^{2q}\cdot\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\xi^{2q},
\end{array}$$

where in (*) we reasoned as for Diracs. Finally it is known that for any integer $q \geq 1$, $\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\xi^{2q} = (2q-1)!!$, where $(2q-1)!! = \prod_{i=1}^q(2i-1) \leq 2^{q-1}q!$. Using (89) we recognize that in both cases

$$\mathbb{E}_{\omega\sim\Lambda}|\langle \pi_0, \phi_\omega\rangle|^{2q}\langle \omega, u\rangle^{2q} \leq \|\pi_0\|_\kappa^2(\varepsilon^2/\sigma^2(s))^q\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\xi^{2q} \leq \|\pi_0\|_\kappa^2(2\varepsilon^2/\sigma^2(s))^q\tfrac{q!}{2}.$$

C.6 Proof of Lemma 3.5

Proof. We exhibit two probability distributions $\tau, \tau' \in \mathfrak{S}^{\mathrm{CT}}(\mathcal{H}_{k,\varepsilon,R})$ such that $\|\tau' - \tau\|_{\Delta\mathcal{L}}/\|\tau' - \tau\|_\kappa$ is bounded from below. Consider $\theta_0 \in \mathbb{R}^d$ with $\|\theta_0\|_2 = 1$ and set $\theta_+ := \frac{\varepsilon}{2}\theta_0, \theta_- = -\theta_+$ (hence $\|\theta_+\| = \|\theta_-\| = \varepsilon/2 \leq R/2$ for small enough ε). Observe that $\varepsilon = \|\theta_+ - \theta_-\|_2$. Setting $\alpha := \frac{R}{2\varepsilon}$, define $h = (c_1, \ldots, c_k)$ with $c_1 = c_+$, $c_l = c_-$ for $l \geq 2$ where

$$c_+ = \theta_+ + \alpha(\theta_+ - \theta_-), \quad c_- = \theta_- + \alpha(\theta_- - \theta_+).$$

Since $\|c_+\| \leq R$, $\|c_-\| \leq R$ and $\|c_+ - c_-\|_2 = (1+\alpha)\|\theta_+ - \theta_-\|_2 = \varepsilon + R/2$ we have $h \in \mathcal{H}_{k,\varepsilon,R}$.

Define two mixtures $\tau = \frac{1}{2}(\delta_{\theta_+} + \delta_{\theta_-})$, $\tau' = \delta_0 \in \mathfrak{S}^{\mathtt{CT}}(\mathcal{H}_{k,\varepsilon,R})$. As $\ell(\theta_+, h) = \ell(\theta_-, h) = (\alpha\varepsilon)^p = (R/2)^p$ and $\ell(0,h) = (1/2+\alpha)^p\varepsilon^p = (R/2)^p(1+\varepsilon/R)^p$, we have

$$\langle \tau' - \tau, \ell(\cdot, h)\rangle = \ (R/2)^p((1+\varepsilon/R)^p - 1). \tag{130}$$

Now set $h' = (c'_1, \ldots, c'_k)$ with $c'_1 = 0$, for all l. Again, $h' \in \mathcal{H}$ and, as $\ell(\theta_+, h') = \ell(\theta_-, h') = (\varepsilon/2)^p$ and $\ell(0,h') = 0$ we have $\langle \tau' - \tau, \ell(\cdot, h')\rangle = \ -(\varepsilon/2)^p$. For $p \in \{1,2\}$ we get

$$\langle \tau' - \tau, \ell(\cdot, h) - \ell(\cdot, h')\rangle = \ (R/2)^p((1+\varepsilon/R)^p - 1) + (\varepsilon/2)^p \geq p(R/2)^p\frac{\varepsilon}{R}. \tag{131}$$

This yields the lower bound $\|\tau - \tau'\|_{\Delta\mathcal{L}(\mathcal{H})} \geq |\langle \tau' - \tau, \ell(\cdot, h) - \ell(\cdot, h')\rangle| \geq p(R/2)^p\frac{\varepsilon}{R}$.

We now upper bound $\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2$. Denote $g(t) := \Phi(t\theta_0)$, $t \in \mathbb{R}$, by assumption the function g is of class $\mathcal{C}^2$ and

$$\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2 = \left\| g(0) - \frac{1}{2}g\Big(-\frac{\varepsilon}{2}\Big) - \frac{1}{2}g\Big(\frac{\varepsilon}{2}\Big)\right\| = \frac{\varepsilon^2}{8}\|g''(0)\| + o(\varepsilon^2),$$

Given (131) we get a constant $c_\Phi > 0$ such that for small enough ε and any $R \geq \varepsilon$, $\|\tau - \tau'\|_{\Delta\mathcal{L}(\mathcal{H})}/\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2 \geq c_\Phi R^{p-1}/\varepsilon$. □

C.7 Link between risks with and without ε-separation

Proof of Lemma 3.8. First, denoting $\eta := d(h,h')$ with $h = (c_1, \ldots, c_k)$, $h' = (c'_1, \ldots, c'_k)$, we show that for any $x \in \mathbb{R}^d$ we have

$$\min_{1\leq j\leq k}\|x - c'_j\| \leq \min_{1\leq i\leq k}\|x - c_i\| + \eta$$

Indeed, denoting $i^\star$ such that $\|x - c_{i^\star}\|_2 = \min_{1\leq i\leq k}\|x - c_i\|$ and $j^\star$ such that $\|c_{i^\star} - c'_{j^\star}\|_2 \leq d(h,h')$, we obtain with the triangle inequality

$$\begin{aligned}\min_{1\leq j\leq k}\|x - c'_j\| \leq \|x - c'_{j^\star}\|_2 &= \|x - c_{i^\star} + c_{i^\star} - c'_{j^\star}\|_2 \\ &\leq \|x - c_{i^\star}\|_2 + \|c_{i^\star} - c'_{j^\star}\|_2 \leq \min_{1\leq i\leq k}\|x - c_i\| + \eta.\end{aligned}$$

For k-medians we obtain $\ell(x,h') \leq \ell(x,h) + \eta$ hence

$$\mathcal{R}_{k-\mathtt{medians}}(\pi, h') = \mathbb{E}_{X\sim\pi}\ell(X,h') \leq \mathbb{E}_{X\sim\pi}\ell(X,h) + \eta \leq \mathcal{R}_{k-\mathtt{medians}}(\pi,h) + \eta.$$

For k-means we have instead

$$\mathcal{R}_{k-\mathtt{means}}(\pi,h') \leq \mathbb{E}_{X\sim\pi}\left(\min_{1\leq i\leq k}\|X - c_i\|_2 + \eta\right)^2 = \mathcal{R}_{k-\mathtt{means}}(\pi,h) + 2\eta\mathcal{R}_{k-\mathtt{medians}}(\pi,h) + \eta^2$$

With Jensen's inequality we have $\mathcal{R}_{k-\mathtt{medians}}(\pi,h) \leq \sqrt{\mathcal{R}_{k-\mathtt{means}}(\pi,h)}$, yielding

$$\sqrt{\mathcal{R}_{k-\mathtt{means}}(\pi,h')} \leq \sqrt{\mathcal{R}_{k-\mathtt{means}}(\pi,h)} + \eta.$$

Exchanging the role of h and h' yields a complementary inequality which completes the proof. □

Proof of Lemma 3.9. Let $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{H}_{k,0,R}$. We proceed by constructing in a greedy way an ε-separated subset of $C_1 := \{c_1, \ldots, c_k\}$ which is also an ε-cover of that set. The construction is standard. Starting at $i = 1$, pick any $c'_i \in C_i$; put $C_{i+1} := C_i \setminus B(c'_i, \varepsilon)$ (where $B(c,\varepsilon)$ denotes the

open ball of center c and radius ε). Iterate until $C_{i+1} = \emptyset$; denote i^* the last iteration. Since the cardinality of C_i is decreasing, we have $i^* \leq k$ iterations. Let $\mathbf{c}' = (c'_1, c'_2, \ldots, c'_{i^*}, c'_{i^*}, \ldots, c'_{i^*})$ (the last element is repeated as needed to attain k centroids). Since $\{c'_1, \ldots, c'_{i^*}\} \subset \{c_1, \ldots, c_k\}$, obviously $d(\mathbf{c}'\|\mathbf{c}) = 0$. On the other hand, by construction $\{c_1, \ldots, c_k\} \subset \bigcup_{1 \leq j \leq i^*} B(c'_j, \varepsilon)$ so that $d(\mathbf{c}\|\mathbf{c}') < \varepsilon$. Finally, also by construction for any $i < i^*$, $c'_{i+1} \notin \bigcup_{1 \leq j \leq i} B(c'_j, \varepsilon)$ and therefore $\mathbf{c}'$ is ε-separated, so that $\mathbf{c}' \in \mathcal{H}_{k,\varepsilon,R}$. Additionally, by the above construction it is clear that any ε-isolated centroid of $\mathbf{c}$ must be selected (once) at some iteration as one of the centroids $c'_j, 1 \leq j \leq i^*$. □

To prove Lemma 3.10 we first establish a refined version of Lemma 3.9.

Lemma C.4. *Given $\varepsilon \geq 0$ and $\mathbf{c} \in \mathcal{H}_{k,0,R}$, there exists $\mathbf{c}' \in \mathcal{H}_{k,\varepsilon,R}$ such that $d(\mathbf{c}, \mathbf{c}') = d(\mathbf{c}, \mathcal{H}_{k,\varepsilon,R})$ and such that all 2ε-isolated centroids of $\mathbf{c}$, $\{c_i, i \in I_{2\varepsilon}(\mathbf{c})\}$ are centroids of $\mathbf{c}'$ (repeated centroids with indices in $I_{2\varepsilon}(\mathbf{c})$ may appear only once in $\mathbf{c}'$.).*

Proof. Let $\mathbf{c}'$ such that $\mathbf{c}' \in \mathcal{H}_{k,\varepsilon,R}$ and $d(\mathbf{c}, \mathbf{c}') = d(\mathbf{c}, \mathcal{H}_{k,\varepsilon,R})$ (the distance $d(\mathbf{c}, \mathcal{H}_{k,\varepsilon,R})$ is attained, since $\mathcal{H}_{k,\varepsilon,R}$ is a compact set). We know by Lemma 3.9 that $d(\mathbf{c}, \mathbf{c}') < \varepsilon$ must hold. Let c_i be any 2ε-isolated centroid of $\mathbf{c}$, and c' a centroid of $\mathbf{c}'$ such that $\|c_i - c'\| < \varepsilon$. By the triangle inequality, for any other centroid $c_j \neq c_i$ of $\mathbf{c}$, $\|c_j - c'\| \geq \|c_j - c_i\| - \|c_i - c'\| > \varepsilon$, and since $d(c_j\|\mathbf{c}') < \varepsilon$, the latter distance is attained for a centroid of $\mathbf{c}'$ different from c'. Hence moving arbitrarily c' can only leave $d(c_j\|\mathbf{c}')$ unaltered or smaller, while obviously also the distance $d(c'_j\|\mathbf{c})$ remains unlaltered for all other centroids $c'_j \neq c'$ of $\mathbf{c}'$ which are unchanged. We can therefore replace c' by c_i in $\mathbf{c}'$ while only making $d(\mathbf{c}\|\mathbf{c}')$, as well as $d(\mathbf{c}'\|\mathbf{c})$, possibly smaller ($d(c'\|\mathbf{c})$ as well as $d(c_i\|\mathbf{c}')$ are set to zero with this operation, the other distances can only shrink by the above argument). We can repeat this operation for all 2ε-isolated centroids of $\mathbf{c}$, leading to the announced claim. □

Proof of Lemma 3.10. Recalling $h^\star = (c_1, \ldots, c_k) \in \mathcal{H}_{k,0,R^\star}$ is the collection of centroids of $\pi^\star = \sum_{i=1}^k \alpha_i \delta_{c_i}$, let $h = (c'_1, \ldots, c'_k)$ be any element in $\mathcal{H}_{k,2\varepsilon,R^\star}$. Denote $I(h^\star, h) := \left\{i; 1 \leq i \leq k : \exists j : c_i = c'_j\right\}$ the index set of centroids of $h^\star$ that are also found in h. Then

$$\mathcal{R}_{\texttt{k-medians}}(\pi^\star, h) = \sum_{i=1}^k \alpha_i \min_{1 \leq j \leq k} \left\|c_i - c'_j\right\| \leq \left(\sum_{i \notin I(h^\star, h)} \alpha_i\right) d(\mathbf{c}\|h).$$

Similarly, $\mathcal{R}_{\texttt{k-means}}(\pi^\star, h) \leq \left(\sum_{i \notin I(h^\star, h)} \alpha_i\right) d(\mathbf{c}\|h)^2$. We apply these estimates to the centroid sets h with the guarantees of Lemma 3.9, resp. C.4 (and take the minimum of the two). Namely, Lemma 3.9 guarantees the existence of $h \in \mathcal{H}_{k,2\varepsilon,R^\star}$ with $I_{2\varepsilon}(h^\star) \subset I(h^\star, h)$ and $d(h^\star, h) \leq 2\varepsilon$, while Lemma C.4 guarantees the existence of $h' \in \mathcal{H}_{k,2\varepsilon,R^\star}$ with $I_{4\varepsilon}(h^\star) \subset I(h^\star, h')$ and $d(h^\star, h') = d(h^\star, \mathcal{H}_{k,2\varepsilon,R^\star})$. Combining with Lemma 3.2 and Lemma 3.8 we obtain the result. □

D Remaining proofs for Sections 3 and 4

D.1 Existence of an (unconstrained) GMM risk minimizer

Let $h^{(n)} = (c_1^{(n)}, ..., c_k^{(n)}, \alpha_1^{(n)}, ..., \alpha_k^{(n)})$ be a sequence such that

$$\mathcal{R}_{\texttt{GMM}}(\pi, h^{(n)}) \overset{n \to \infty}{\longrightarrow} \mathcal{R}^*_{\texttt{GMM}} := \inf_{h \in \mathbb{R}^{kd} \times \mathbb{S}_{k-1}} \mathcal{R}_{\texttt{GMM}}(\pi, h).$$

By continuity of $h \mapsto \mathcal{R}_{\texttt{GMM}}(\pi, h)$, it suffices to prove that $(h^{(n)})_{n \geq 1}$ has an accumulation point in order to establish the existence of a minimum of $\mathcal{R}_{\texttt{GMM}}(\pi, h)$. If all the centroids $c_i^{(n)}$ remain bounded, this

is the case, by compactness. If not, we can assume without loss of generality (up to permutation and taking a subsequence) that $\left\|c_1^{(n)}\right\| \to \infty$.

Let $h_0^{(n)} := (0, c_2^{(n)}, \dots, c_k^{(n)}, \alpha_1^{(n)}, \dots, \alpha_k^{(n)})$. Then, denoting, for any $c \in \mathbb{R}^d$, $\phi_c(t) := \exp(-\|t\|_{\mathbf{\Sigma}}^2/2)$ the unnormalized density of the Gaussian distribution centered in c and of covariance $\mathbf{\Sigma}$, it holds for any input point x:

$$\begin{aligned}
-\log \pi_{h_0^{(n)}}(x) - (-\log \pi_{h^{(n)}}(x)) &= \log\left(\frac{\sum_{i=1}^k \alpha_i^{(n)} \phi_{c_i^{(n)}}(x)}{\alpha_1^{(n)}\phi_0(x) + \sum_{j=2}^k \alpha_j^{(n)} \phi_{c_j^{(n)}}(x)}\right) \\
&= \log\left(1 + \frac{\alpha_1^{(n)}(\phi_{c_1^{(n)}}(x) - \phi_0(x))}{\alpha_1^{(n)}\phi_0(x) + \sum_{j=2}^k \alpha_j^{(n)} \phi_{c_j}^{(n)}(x)}\right) \\
&\leq \log\left(1 + \left(\frac{\phi_{c_1^{(n)}}(x)}{\phi_0(x)} - 1\right)_+\right) \\
&= \left(\log\left(\frac{\phi_{c_1^{(n)}}(x)}{\phi_0(x)}\right)\right)_+ \\
&= \frac{1}{2}\left(\|x\|_{\mathbf{\Sigma}}^2 - \left\|x - c_1^{(n)}\right\|_{\mathbf{\Sigma}}^2\right)_+ \\
&= \left(\left\langle x, c_1^{(n)}\right\rangle_{\mathbf{\Sigma}} - \frac{1}{2}\left\|c_1^{(n)}\right\|_{\mathbf{\Sigma}}^2\right)_+ .
\end{aligned}$$

Taking expectations (note that integrability follows from the existence of the first moment of π, itself following from the assumption of GMM loss integrability under π, which implies existence of moments up to order 2),

$$\mathcal{R}_{\texttt{GMM}}(\pi, h_0^{(n)}) - \mathcal{R}_{\texttt{GMM}}(\pi, h^{(n)}) \leq \mathbb{E}_{X\sim\pi}\left[\left(\left\langle X, c_1^{(n)}\right\rangle_{\mathbf{\Sigma}} - \frac{1}{2}\left\|c_1^{(n)}\right\|_{\mathbf{\Sigma}}^2\right)_+\right].$$

Let $f_c : x \mapsto \left(\langle x, c\rangle_{\mathbf{\Sigma}} - \frac{1}{2}\|c\|_{\mathbf{\Sigma}}^2\right)_+$; we have that $\sup_c f_c(x) = \frac{1}{2}\|x\|_{\mathbf{\Sigma}}^2$, and f_c converges pointwise to 0 as $\|c\|_{\mathbf{\Sigma}} \to \infty$. Since $\left\|c_1^{(n)}\right\|_{\mathbf{\Sigma}} \to \infty$, and π has finite second order moments, by dominated convergence we get that the right-hand side above converges to 0, and that $\lim_{n\to\infty} \mathcal{R}_{\texttt{GMM}}(\pi, h_0^{(n)}) = \lim_{n\to\infty} \mathcal{R}_{\texttt{GMM}}(\pi, h^{(n)}) = \mathcal{R}^*_{\texttt{GMM}}$. Repeating this operation as necessary with other centroids diverging to infinity, we see that we can replace the sequence $h^{(n)}$ by a sequence remaining in a compact and with the same limit for the risk, for which an accumulation point exists, attaining the minimum of the risk.

D.2 Control of $\|\mathcal{D}\|_{\Delta\mathcal{L}}$

For compressive k-means / k-medians, with the loss defined in (17), we consider a constrained hypothesis class $\mathcal{H}$ such that $\mathfrak{S}^\star(\mathcal{H}) = \mathfrak{S}_k(\mathcal{T})$, where $\mathcal{T} := \mathcal{T}_{\texttt{Dirac}} = (\mathbf{\Theta}_R, \|\cdot\|_2/\varepsilon, \varphi)$ is as in Definition 6.9, depending on some separation parameter and radius $0 < \varepsilon \leq R$; we recall $\mathbf{\Theta}_R = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R)$ and $\varphi(\theta) = \delta_\theta$, and underline that while the separation parameter ε does not change the base distribution set $\varphi(\mathbf{\Theta})$, it will determine separation in the mixture and dipole sets derived from it.

Lemma D.1. *Consider $0 < \varepsilon \leq R$, $\mathcal{T} = \mathcal{T}_{\texttt{Dirac}}$ based on the parameter set $\mathbf{\Theta}_R = \big\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\big\}$. Consider $\mathcal{L}(\mathcal{H})$ associated to k-means (resp. k-medians) with $\mathcal{H} \subseteq \mathcal{H}_R := \{h = (c_1, \dots, c_k), \|c_l\|_2 \leq R\}$. For any shift-invariant kernel κ that is 1-strongly locally characteristic with respect to $\mathcal{T}$ we have*

$$\|\mathcal{D}\|_{\Delta\mathcal{L}} \leq 2 \cdot \|\pi_0\|_\kappa^{-1} \cdot (2R)^p \tag{132}$$

with $p = 2$ for k-means, $p = 1$ for k-medians.

For compressive Gaussian mixture modeling we use $\mathcal{T} := \mathcal{T}_{\texttt{Gauss}}$ as in Definition 6.9.

Lemma D.2. *Consider $0 < \varepsilon \leq R$, $\mathcal{T} = \mathcal{T}_{\texttt{Gauss}}$ based on $\mathbf{\Theta}_R = \big\{\theta \in \mathbb{R}^d : \|\theta\|_{\mathbf{\Sigma}} \leq R\big\}$, and $\mathcal{L}(\mathcal{H})$ associated to k-mixtures of[10] Gaussians $\pi_l = \mathcal{N}(c_l, \mathbf{\Sigma})$, with $\mathcal{H} \subseteq \mathcal{H}_R := \{h = (c_1, \dots, c_k, \boldsymbol{\alpha}), \|c_l\|_{\mathbf{\Sigma}} \leq R, \boldsymbol{\alpha} \in \mathbb{S}_{k-1}\}$.*

For any shift-invariant kernel κ that is 1-strongly locally characteristic with respect to $\mathcal{T}$ we have

$$\|\mathcal{D}\|_{\Delta\mathcal{L}} \leq 2 \cdot \|\pi_0\|_\kappa^{-1} \cdot 2R^2 \tag{133}$$

Proof of Lemmas-D.1-D.2. Since κ is 1-strongly locally characteristic with respect to $\mathcal{T}$ we can use Theorem 5.15. Slightly abusing notation (confusing $\mathcal{T} = (\mathbf{\Theta}, \varrho, \varphi)$ with $\varphi(\mathbf{\Theta})$) we denote $\|\mathcal{T}\|_{\mathcal{G}} := \|\varphi(\mathbf{\Theta})\|_{\mathcal{G}}$ and observe that $\|\mathcal{M}\|_{\mathcal{G}} = \|\pi_0\|_\kappa^{-1} \cdot \|\mathcal{T}\|_{\mathcal{G}}$, and that $\varphi : \theta \mapsto \pi_\theta$ is $L'_{\mathcal{G}}$-Lipschitz with respect to $\|\cdot\|$, $\|\cdot\|_{\mathcal{G}}$ if, and only if ψ is $L_{\mathcal{G}}$-Lipschitz with respect to $\|\cdot\|$ and $\|\cdot\|_{\mathcal{G}}$, with $L_{\mathcal{G}} = L'_{\mathcal{G}}\|\pi_0\|_\kappa^{-1}$. By Theorem 5.15, if we can show that $\varphi : \theta \mapsto \pi_\theta$ is $L'_{\Delta\mathcal{L}}$-Lipschitz with respect to $\|\cdot\|$, $\|\cdot\|_{\Delta\mathcal{L}}$ then ψ has the desired Lipschitz property with $L_{\Delta\mathcal{L}} \leq \|\pi_0\|_\kappa^{-1} L'_{\Delta\mathcal{L}}$ and

$$\|\pi_0\|_\kappa^{-1} \cdot \|\mathcal{T}\|_{\Delta\mathcal{L}} = \|\mathcal{M}\|_{\Delta\mathcal{L}} \leq \|\mathcal{D}\|_{\Delta\mathcal{L}} \leq \|\pi_0\|_\kappa^{-1} \big(L'_{\mathcal{G}} + \|\mathcal{T}\|_{\Delta\mathcal{L}}\big).$$

The rest of the proof consists in characterizing $\|\mathcal{T}\|_{\Delta\mathcal{L}}$ and bounding $L'_{\Delta\mathcal{L}}$.

For this we consider $\Delta\ell(\cdot, h, h') = \ell(\cdot, h) - \ell(\cdot, h') \in \Delta\mathcal{L}(\mathcal{H}_R)$ where $h, h' \in \mathcal{H}_R$.

With $\mathcal{T} = \mathcal{T}_{\texttt{Dirac}}$ and the loss associated to compressive clustering, given $h = (c_1, \dots, c_k)$, for each $\theta \in \mathbf{\Theta}_R$ the triangle inequality yields $\|\theta - c_l\|_2 \leq \|\theta\|_2 + \|c_l\|_2 \leq 2R$ hence $0 \leq \ell(\theta, h) \leq (2R)^p$ where we recall that $p = 2$ for k-means and $p = 1$ for k-medians. Similarly $0 \leq \ell(\theta, h') \leq (2R)^p$ hence $g(\theta) := \mathbb{E}_{X \sim \pi_\theta} \Delta\ell(X, h, h') = \Delta\ell(\theta, h, h')$ satisfies $g(\theta) \leq (2R)^p$. This shows that

$$\|\mathcal{T}_{\texttt{Dirac}}\|_{\Delta\mathcal{L}} \leq (2R)^p.$$

The bound is reached using θ such that $\|\theta\|_2 = R$, $c_1 = \ldots = c_k = -\theta$, $h = (c_1, \dots, c_k)$, $h' = -h$.

Given $\theta, \theta' \in \mathbf{\Theta}_R$, let i be an index such that $\ell(\theta', h) = \min_l \|\theta' - c_l\|_2^p = \|\theta' - c_i\|_2^p$. By definition, $\ell(\theta, h) = \min_l \|\theta - c_l\|_2^p \leq \|\theta - c_i\|_2^p$ hence $\ell(\theta, h) - \ell(\theta', h) \leq \|\theta - c_i\|_2^p - \|\theta' - c_i\|_2^p$. Similarly, with j such that $\ell(\theta, h') = \|\theta - c'_j\|_2^p$ (where $h' = (c'_1, \dots, c'_k)$) we get $\ell(\theta', h') - \ell(\theta, h') \leq \big\|\theta' - c'_j\big\|_2^p - \big\|\theta - c'_j\big\|_2^p$ hence

$$\begin{aligned} g(\theta) - g(\theta') = [\ell(\theta, h) - \ell(\theta, h')] - [\ell(\theta', h) - \ell(\theta', h')] &= [\ell(\theta, h) - \ell(\theta', h)] + [\ell(\theta', h') - \ell(\theta, h')] \\ &\leq \|\theta - c_i\|_2^p - \|\theta' - c_i\|_2^p + \big\|\theta' - c'_j\big\|_2^p - \big\|\theta - c'_j\big\|_2^p. \end{aligned}$$

For k-medians, $p = 1$ and the reversed triangle inequality further yields

$$\|\theta - c_i\|_2 - \|\theta' - c_i\|_2 + \big\|\theta' - c'_j\big\|_2 - \big\|\theta - c'_j\big\|_2 \leq 2\|\theta - \theta'\|_2 = 2(2R)^{p-1}\|\theta - \theta'\|_2.$$

In the case of k-means, $p = 2$ and we use

$$\begin{aligned} \|\theta - c_i\|_2^2 - \|\theta' - c_i\|_2^2 + \big\|\theta' - c'_j\big\|_2^2 - \big\|\theta - c'_j\big\|_2^2 &= 2\big\langle \theta - \theta', c'_j - c_i \big\rangle \\ &\leq 2\|\theta - \theta'\|_2 \cdot \big\|c'_j - c_i\big\|_2 \\ &\leq 4R\|\theta - \theta'\|_2 = 2(2R)^{p-1}\|\theta - \theta'\|_2. \end{aligned}$$

[10] not-necessarily separated

By symmetry we obtain $|g(\theta) - g(\theta')| \leq 2(2R)^{p-1}\|\theta - \theta'\|_2$. As this holds for any $\theta, \theta' \in \mathbf{\Theta}_R$ and $g \in \Delta\mathcal{L}$, and as $\|\theta - \theta'\|_2 = \varepsilon\|\theta - \theta'\|$, we get $L'_{\Delta\mathcal{L}} \leq \varepsilon 2(2R)^{p-1} = (\varepsilon/R)\|\mathcal{T}_{\texttt{Dirac}}\|_{\Delta\mathcal{L}}$.

With $\mathcal{T} = \mathcal{T}_{\texttt{Gauss}}$ and the loss associated to compressive GMM, we prove at the end of this section that for any $h, h' \in \mathcal{H}_R$ the function $g(\theta) := \mathbb{E}_{X\sim\pi_\theta}\Delta\ell(X, h, h')$ satisfies

$$|g(\theta)| \leq 2R^2 \tag{134}$$

$$|g(\theta) - g(\theta')| \leq 2R\varepsilon\|\theta - \theta'\|. \tag{135}$$

where the first bound is reached. We obtain $\|\mathcal{T}_{\texttt{Gauss}}\|_{\Delta\mathcal{L}} = 2R^2$, $L'_{\Delta\mathcal{L}} \leq 2\varepsilon R = (\varepsilon/R)\|\mathcal{T}_{\texttt{Gauss}}\|_{\Delta\mathcal{L}}$.

In both cases since $R \geq \varepsilon$ we have $L'_{\Delta\mathcal{L}} + \|\mathcal{T}\|_{\Delta\mathcal{L}} \leq 2\|\mathcal{T}\|_{\Delta\mathcal{L}}$. □

The following lemma, which applies to any family of absolutely continuous probability distributions $\{\pi_\theta : \theta \in \mathbf{\Theta}\}$ on $\mathbb{R}^d$, will be soon specialized to Gaussians with fixed known covariance.

Lemma D.3. *Consider a family of probability distributions $\{\pi_\theta : \theta \in \mathbf{\Theta}\}$ on $\mathbb{R}^d$ having a density with respect to the Lebesgue measure. Recalling* H *denotes the differential entropy* (33)*, assume that*

$$\mathrm{H}_{\min} := \inf_{\theta\in\mathbf{\Theta}} \mathrm{H}(\pi_\theta) > -\infty; \tag{136}$$

$$\mathrm{H}_{\max} := \sup_{\theta,\theta'\in\mathbf{\Theta}} \mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta||\pi_{\theta'}) < \infty. \tag{137}$$

For any $\pi_h := \sum_{l=1}^k \alpha_l \pi_{\theta_l}$, where $\boldsymbol{\alpha} \in \mathbb{S}_{k-1}$, $\theta_l \in \mathbf{\Theta}$ we have for any $\theta \in \mathbf{\Theta}$:

$$\mathrm{H}_{\min} \leq \mathbb{E}_{X\sim\pi_\theta}[-\log \pi_h(X)] \leq \mathrm{H}_{\max}.$$

The lower and the upper bounds are both tight.

Proof. By the definition of the Kullback-Leibler divergence and its convexity properties, we have

$$\begin{aligned}\mathrm{H}(\pi_\theta) \leq \mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta||\pi_h) &= \mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta||\sum_{l=1}^k \alpha_l \pi_{\theta_l}) \\ &\leq \mathrm{H}(\pi_\theta) + \sum_{l=1}^k \alpha_l \mathrm{KL}(\pi_\theta||\pi_{\theta_l}) \leq \mathrm{H}(\pi_\theta) + \sup_{\theta'\in\mathbf{\Theta}} \mathrm{KL}(\pi_\theta||\pi_{\theta'}).\end{aligned}$$

For a given θ, both the lower and the upper bound are tight. The conclusion immediately follows. □

This translates into a concrete result for Gaussian mixtures with fixed known covariance.

Proof of Equations (134)-(135) *- end of the proof of Lemma D.2.* To establish (134) we exploit Lemma D.3. The entropy of a Gaussian is $\mathrm{H}(\pi_\theta) = \frac{1}{2}\log\det(2\pi e\mathbf{\Sigma})$ which is independent of θ, hence $\mathrm{H}_{\min} = \frac{1}{2}\log\det(2\pi e\mathbf{\Sigma})$. The Kullback-Leibler divergence has a closed form expression in the case of multivariate Gaussians (see e.g. Duchi, 2007):

$$\mathrm{KL}(\mathcal{N}(\theta_1, \mathbf{\Sigma}_1)||\mathcal{N}(\theta_2, \mathbf{\Sigma}_2)) = \tfrac{1}{2}\left[\log\frac{\det(\mathbf{\Sigma}_2)}{\det(\mathbf{\Sigma}_1)} + \mathrm{tr}\big(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\big) - d + (\theta_2 - \theta_1)^T\mathbf{\Sigma}_2^{-1}(\theta_2 - \theta_1)\right]; \tag{138}$$

hence $\mathrm{KL}(\pi_\theta||\pi_{\theta'}) = \frac{1}{2}\|\theta - \theta'\|^2_{\mathbf{\Sigma}}$. Since $\|\theta\|_{\mathbf{\Sigma}} \leq R$ when $\theta \in \mathbf{\Theta}_R$, we have $\mathrm{H}_{\max} = \mathrm{H}_{\min} + \frac{1}{2}\|\theta - \theta'\|^2_{\mathbf{\Sigma}} \leq \mathrm{H}_{\min} + 2R^2$. By Lemma D.3 we obtain (134) as follows

$$|g(\theta)| = |\mathbb{E}_{X\sim\pi_\theta}\Delta\ell(X, h, h')| = |\mathbb{E}_{X\sim\pi_\theta}[-\log\pi_h(X)] - \mathbb{E}_{X\sim\pi_\theta}[-\log\pi_{h'}(X)]| \leq \mathrm{H}_{\max} - \mathrm{H}_{\min} \leq 2R^2.$$

where the bound is tight.

We now turn to Equation (135). Denoting $\ell_h(x) := -\log \pi_h(x)$ and $f_h(\theta) := \mathbb{E}_{X\sim\pi_\theta}\ell_h(X)$, since $\pi_\theta(x) = \pi_0(x-\theta) = \pi_0(\theta - x)$ we have

$$f_h(\theta) = \int \pi_\theta(x)\ell_h(x)dx = \int \pi_0(\theta-x)\ell_h(x)dx = \int \pi_0(x)\ell_h(\theta-x)dx;$$

$$\nabla f_h(\theta) = \int \pi_0(x)\nabla\ell_h(\theta-x)dx = \int \pi_0(\theta-x)\nabla\ell_h(x)dx = \mathbb{E}_{X\sim\pi_\theta}\nabla\ell_h(X);$$

$$\nabla\ell_h(x) = -\frac{\nabla\pi_h(x)}{\pi_h(x)} = -\frac{\sum_l \alpha_l\pi_{\theta_l}(x)\cdot\frac{\nabla\pi_{\theta_l}(x)}{\pi_{\theta_l}(x)}}{\pi_h(x)} = -\sum_l \frac{\alpha_l\pi_{\theta_l}(x)}{\pi_h(x)}\cdot\frac{\nabla\pi_{\theta_l}(x)}{\pi_{\theta_l}(x)} = -\sum_l \beta_l(x)\cdot\nabla\log\pi_{\theta_l}(x),$$

where $\beta_l(x) := \frac{\alpha_l\pi_{\theta_l}(x)}{\sum_l \alpha_l\pi_{\theta_l}(x)} \geq 0$ satisfies $\sum_l \beta_l(x) = 1$. Since $\nabla\log\pi_{\theta_l}(x) = -\mathbf{\Sigma}^{-1}(x-\theta_l)$, we obtain

$$\nabla f_h(\theta) = \mathbb{E}_{X\sim\pi_\theta}\sum_l \beta_l(X)\cdot\mathbf{\Sigma}^{-1}(X-\theta_l) = \mathbf{\Sigma}^{-1}\mathbb{E}_{X\sim\pi_\theta}\big(X - \sum_l \beta_l(X)\theta_l\big) = \mathbf{\Sigma}^{-1}\big(\theta - \sum_l \gamma_l\cdot\theta_l\big),$$

with $\gamma_l := \mathbb{E}_{X\sim\pi_\theta}\beta_l(X) \geq 0$, $\sum_l \gamma_l = 1$. Similarly we have $\nabla f_{h'}(\theta) = \mathbf{\Sigma}^{-1}(\theta - \sum_l \gamma'_l\cdot\theta'_l)$ where $\gamma'_l \geq 0$ and $\sum_l \gamma'_l = 1$. Since $g(\theta) = f_h(\theta) - f_{h'}(\theta)$, and $\|\theta_l\|_{\mathbf{\Sigma}} \leq R$, $\|\theta'_l\|_{\mathbf{\Sigma}} \leq R$, we get

$$\begin{aligned}\|\nabla g(\theta)\|_{\mathbf{\Sigma}^{-1}} = \Big\|\mathbf{\Sigma}^{1/2}(\nabla f_h(\theta) - \nabla f_{h'}(\theta))\Big\|_2 &= \Big\|\mathbf{\Sigma}^{-1/2}\big(\sum_l \gamma'_l\cdot\theta'_l - \sum_l \gamma_l\cdot\theta_l\big)\Big\|_2 \\ &= \Big\|\sum_l \gamma'_l\cdot\theta'_l - \sum_l \gamma_l\cdot\theta_l\Big\|_{\mathbf{\Sigma}} \leq 2R.\end{aligned}$$

To obtain (135), given θ, θ', defining $\theta(t) := \theta + t(\theta' - \theta)$ we have

$$\begin{aligned}|g(\theta') - g(\theta)| &= \left|\int_0^1 \frac{d}{dt}g(\theta(t))dt\right| = \left|\int_0^1 \langle\nabla g(\theta(t)), \theta' - \theta\rangle dt\right| \leq \int_0^1 \|\nabla g(\theta(t))\|_{\mathbf{\Sigma}^{-1}}\|\theta' - \theta\|_{\mathbf{\Sigma}}dt \\ &\leq 2R\|\theta' - \theta\|_{\mathbf{\Sigma}} = 2R\varepsilon\|\theta - \theta'\|.\end{aligned}$$

□

D.3 Elements to control the bias term

While Gaussian Mixture Modeling is a maximum likelihood task, with a negative log-likelihood loss, both k-means and k-medians are *compression-type tasks* as defined in [Gribonval et al., 2021]:

Definition D.4 (Compression-type task, [Gribonval et al., 2021, Definition 3.1])**.** *We call the learning task a* compression-type task *if the loss can be written as $\ell(x,h) = d^p(x, P_h x)$, where d is a metric on the sample space $\mathcal{Z}$, $p > 0$, and $P_h : \mathcal{Z} \to \mathcal{Z}$ is a projection function, i.e., $P_h \circ P_h = P_h$ and $d(x, P_h x) \leq d(x, P_h x')$ for all $x, x' \in \mathcal{Z}$.*

For k-means and k-medians, $d(x,x') = \|x - x'\|_2$ is the Euclidean distance on $\mathcal{Z} = \mathbb{R}^d$. Given $h = (c_1, \ldots, c_k)$, the function P_h maps x to the closest c_j, with ties broken arbitrarily, and can be used to define a *Voronoi partition* $W_j(h) := P_h^{-1}(c_j) = \big\{x \in \mathbb{R}^d : P_h x = c_j\big\}$, i.e. a collection of pairwise disjoint sets such that $\cup_j W_j(h) = \mathbb{R}^d$ and $W_j(h) \subseteq V_j(h)$ with $V_j(h)$ the Voronoi cells defined in (18). The push-forward $P_h\pi$ of a probability distribution π through P_h is the probability distribution of $Y = P_h X$ when $X \sim \pi$. Here it reads more explicitly as $P_h\pi = \sum_{j=1}^k \alpha_j\delta_{c_j}$ with $\alpha_j = \pi(X \in W_j(h))$.

The goal of the next result is to have a device to relate the excess risk with respect to hypotheses in the restricted class $\mathcal{H}$, which will be controlled via Theorem 2.2 (15)-(16), to the excess risk with

respect to the optimal in an unconstrained (or less constrained) class, $\overline{\mathcal{H}}$, e.g. $\overline{\mathcal{H}} = (\mathbb{R}^d)^k$ for k-means (resp. $\overline{\mathcal{H}} = (\mathbb{R}^d)^k \times \mathbb{S}_{k-1}$ for GMM). Observe that the main control (15) on the restricted hypothesis class takes the form

$$\forall h_0 \in \mathcal{H} : \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) \leq d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S}) + \Delta(\pi, \hat{\pi}_n),$$

where the trailing rest term $\Delta(\pi, \hat{\pi}_n)$ does not depend on h_0. If $h^\star$ denotes the optimal hypothesis over the larger class $\overline{\mathcal{H}}$, we deduce from the above:

$$\forall h_0 \in \mathcal{H} : \Delta\mathcal{R}_{h^\star}(\pi, \hat{h}) = \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) + \Delta\mathcal{R}_{h^\star}(\pi, h_0) \leq \big(\Delta\mathcal{R}_{h^\star}(\pi, h_0) + d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S})\big) + \Delta(\pi, \hat{\pi}_n).$$

It is therefore of interest to further upper bound the first term in the above estimate. This is what we obtain in the following result.

Lemma D.5. *Consider a compression-type task with P_h defined for any $h \in \overline{\mathcal{H}}$. With the notations and assumptions of Theorem 2.2 on a class $\mathcal{H} \subseteq \overline{\mathcal{H}}$ with $\mathfrak{S} = \mathfrak{S}^\star(\mathcal{H})$, considering π a probability distribution on $\mathcal{Z}$ with integrable loss, $h^\star \in \arg\min_{h \in \overline{\mathcal{H}}} \mathcal{R}(\pi, h)$, and $\pi^\star := P_{h^\star}\pi$ we have*

$$\inf_{h_0 \in \mathcal{H}} \big\{\Delta\mathcal{R}_{h^\star}(\pi, h_0) + d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S})\big\} \leq D^{\overline{\mathcal{H}}}_{h^\star}(\pi \| \pi^\star) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 + d^{\mathcal{H}}(\pi^\star, \mathfrak{S}),$$

with

$$d^{\mathcal{H}}(\pi^\star, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \left\{ \sup_{h \in \mathcal{H}} (\mathcal{R}(\pi^\star, h) - \mathcal{R}(\tau, h)) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 \right\}. \tag{139}$$

The same holds for a maximum likelihood task with $\ell(x, h) = -\log \pi_h(x)$, using $\mathfrak{S} = \mathfrak{S}^{\mathtt{ML}}(\mathcal{H})$, $\pi^\star = \pi_{h^\star}$ and

$$d^{\mathcal{H}}(\pi^\star, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \left\{ \sup_{h \in \mathcal{H}} (\mathrm{KL}(\pi^\star || \pi_h) - \mathrm{KL}(\tau || \pi_h)) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 \right\}. \tag{140}$$

Proof. Let $h_0, h \in \mathcal{H}$, and $\tau \in \mathfrak{S}$ be given. Since $\Delta\mathcal{R}_a(\pi, b) + \Delta\mathcal{R}_b(\pi, c) = \Delta\mathcal{R}_a(\pi, c)$ for $a, b, c \in \overline{\mathcal{H}}$, we have

$$\begin{aligned}
\Delta\mathcal{R}_{h^\star}(\pi, h_0) + \Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(\tau, h) =& \Delta\mathcal{R}_{h^\star}(\pi, h) - \Delta\mathcal{R}_{h^\star}(\tau, h) - \Delta\mathcal{R}_{h_0}(\tau, h^\star) \\
=& \Delta\mathcal{R}_{h^\star}(\pi, h) - \Delta\mathcal{R}_{h^\star}(\tau, h) + \Delta\mathcal{R}_{h^\star}(\tau, h_0) \\
=& (\Delta\mathcal{R}_{h^\star}(\pi, h) - \Delta\mathcal{R}_{h^\star}(\pi^\star, h)) \\
& + (\Delta\mathcal{R}_{h^\star}(\pi^\star, h) - \Delta\mathcal{R}_{h^\star}(\tau, h)) + \Delta\mathcal{R}_{h^\star}(\tau, h_0).
\end{aligned}$$

Taking the supremum over $h \in \mathcal{H} \subseteq \overline{\mathcal{H}}$ and denoting $D^{\mathcal{H}}_{h^\star}(\pi^\star \| \tau) := \sup_{h \in \mathcal{H}} (\Delta\mathcal{R}_{h^\star}(\pi^\star, h) - \Delta\mathcal{R}_{h^\star}(\tau, h))$ (even though $h^\star$ may not belong to $\mathcal{H}$) yields

$$\Delta\mathcal{R}_{h^\star}(\pi, h_0) + D^{\mathcal{H}}_{h_0}(\pi \| \tau) \leq D^{\overline{\mathcal{H}}}_{h^\star}(\pi \| \pi^\star) + D^{\mathcal{H}}_{h^\star}(\pi^\star \| \tau) + \Delta\mathcal{R}_{h^\star}(\tau, h_0),$$

hence by a triangle inequality

$$\begin{aligned}
\Delta\mathcal{R}_{h^\star}(\pi, h_0) + D^{\mathcal{H}}_{h_0}(\pi \| \tau) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2 \leq& D^{\overline{\mathcal{H}}}_{h^\star}(\pi \| \pi^\star) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 \\
& + D^{\mathcal{H}}_{h^\star}(\pi^\star \| \tau) + \Delta\mathcal{R}_{h^\star}(\tau, h_0) \\
& + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2.
\end{aligned} \tag{141}$$

The joint infimum of (141) over $h_0 \in \mathcal{H}$ and $\tau \in \mathfrak{S}$ yields

$$\begin{aligned}
\inf_{h_0 \in \mathcal{H}} \Big\{\Delta\mathcal{R}_{h^\star}(\pi, h_0) + d^{\mathcal{H}}_{h_0}(\pi, \mathfrak{S})\Big\} \leq& D^{\overline{\mathcal{H}}}_{h^\star}(\pi \| \pi^\star) + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 \\
& + \inf_{\tau \in \mathfrak{S}} \left\{ \Big[D^{\mathcal{H}}_{h^\star}(\pi^\star \| \tau) + \inf_{h_0 \in \mathcal{H}} \Delta\mathcal{R}_{h^\star}(\tau, h_0)\Big] + (2+\nu)C_{\mathcal{A}} \|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 \right\}.
\end{aligned} \tag{142}$$

For any $h \in \overline{\mathcal{H}}$ we have $\Delta\mathcal{R}_{h^\star}(\pi^\star, h) - \Delta\mathcal{R}_{h^\star}(\tau, h) = \mathcal{R}(\pi_\star, h) - \mathcal{R}(\tau, h) + \mathcal{R}(\tau, h^\star) - \mathcal{R}(\pi^\star, h^\star)$ hence $D^{\mathcal{H}}_{h^\star}(\pi^\star \| \tau) = \sup_{h\in\mathcal{H}}(\mathcal{R}(\pi_\star, h) - \mathcal{R}(\tau, h)) + \mathcal{R}(\tau, h^\star) - \mathcal{R}(\pi^\star, h^\star)$.

To conclude observe that for a compression-type task, we have $\mathcal{R}(\pi^\star, h^\star) = 0$ and, for $\tau \in \mathfrak{S}^\star(\mathcal{H})$,

$$\inf_{h_0\in\mathcal{H}} \Delta\mathcal{R}_{h^\star}(\tau, h_0) = \inf_{h_0\in\mathcal{H}} \{\mathcal{R}(\tau, h_0) - \mathcal{R}(\tau, h^\star)\} = -\mathcal{R}(\tau, h^\star).$$

As a result $D^{\mathcal{H}}_{h^\star}(\pi^\star\|\tau) + \inf_{h_0\in\mathcal{H}} \Delta\mathcal{R}_{h^\star}(\tau, h_0) = \sup_{h\in\mathcal{H}}(\mathcal{R}(\pi_\star, h) - \mathcal{R}(\tau, h))$.

For a maximum likelihood task, $\mathcal{R}(\pi^\star, h^\star) = \mathrm{H}(\pi^\star)$ with $H(\cdot)$ the entropy, and $\mathcal{R}(\tau, h_0) = \mathrm{KL}(\tau||\pi_{h_0}) + \mathrm{H}(\tau)$ for $h_0 \in \mathcal{H}$, hence

$$\inf_{h_0\in\mathcal{H}} \Delta\mathcal{R}_{h^\star}(\tau, h_0) = \inf_{h_0\in\mathcal{H}} \{\mathcal{R}(\tau, h_0) - \mathcal{R}(\tau, h^\star)\} = \mathrm{H}(\tau) - \mathcal{R}(\tau, h^\star);$$

as a result

$$\begin{aligned} D^{\mathcal{H}}_{h^\star}(\pi^\star\|\tau) + \inf_{h_0\in\mathcal{H}} \Delta\mathcal{R}_{h^\star}(\tau, h_0) &= \sup_{h\in\mathcal{H}}(\mathcal{R}(\pi_\star, h) - \mathcal{R}(\tau, h)) + \mathrm{H}(\tau) - \mathrm{H}(\pi^\star) \\ &= \sup_{h\in\mathcal{H}}(\mathrm{KL}(\pi^\star||\pi_h) - \mathrm{KL}(\tau||\pi_h)). \end{aligned}$$

□

Next we deal with the term $D^{\overline{\mathcal{H}}}_{h^\star}(\pi\|\pi^\star)$ in Lemma D.5. For k-medians by [Gribonval et al., 2021, Lemma 3.2] we have $D^{\overline{\mathcal{H}}}_{h^\star}(\pi\|\pi^\star) = 0$. We now show that this also holds for k-means when $\overline{\mathcal{H}} = (\mathbb{R}^d)^k$.

Lemma D.6. *Consider ℓ the loss associated to k-means on a class $\mathcal{H}$ and π a probability distribution on $\mathbb{R}^d$ with integrable loss.*

- $D^{\mathcal{H}}_{h_0}(\pi\|P_{h_0}\pi) = 0$ *for each* $h_0 = (c_1, \dots, c_k) \in \mathcal{H}$ *such that*

$$\pi(X \in W_j(h_0)) \neq 0 \Longrightarrow c_j = \mathbb{E}_\pi(X|X \in W_j(h_0)), \qquad \forall 1 \leq j \leq k, \tag{143}$$

- $D^{\overline{\mathcal{H}}}_{h^\star}(\pi\|P_{h^\star}\pi) = 0$ *for each* $h^\star \in \arg\min_{h\in\overline{\mathcal{H}}} \mathcal{R}(\pi, h)$ *with* $\overline{\mathcal{H}} = (\mathbb{R}^d)^k$.
- *If* $\mathcal{H} \subseteq \mathcal{H}_R := \{h = (c_1, \dots, c_k), \|c_l\|_2 \leq R\}$ *then*

$$D^{\mathcal{H}}_{h_0}(\pi\|P_{h_0}\pi) \leq 4R \cdot \mathcal{R}_{\texttt{k-medians}}(\pi, h_0), \qquad \forall h_0 \in \mathcal{H}. \tag{144}$$

Proof. By [Gribonval et al., 2021, Equation (30)], for any $x \in \mathcal{Z} = \mathbb{R}^d$ and $h \in \mathcal{H}$ we have

$$\|x - P_h x\|_2^2 \leq \|x - P_h P_{h_0} x\|_2^2 = \|x - P_{h_0} x\|_2^2 + \|P_{h_0} x - P_h P_{h_0} x\|_2^2 + 2\langle x - P_{h_0} x, P_{h_0} x - P_h P_{h_0} x\rangle.$$

It follows thus from [Gribonval et al., 2021, Equation (68)] that

$$\begin{aligned} \Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(P_{h_0}\pi, h) &\leq \mathbb{E}_{X\sim\pi}\big\{\|X - P_h X\|_2^2 - \|X - P_{h_0} X\|_2^2 - \|P_{h_0} X - P_h P_{h_0} X\|_2^2\big\} \\ &\leq 2\mathbb{E}_{X\sim\pi}\langle X - P_{h_0} X, P_{h_0} X - P_h P_{h_0} X\rangle. \end{aligned}$$

We have $\langle x - P_{h_0} x, P_{h_0} x - P_h P_{h_0} x\rangle = \sum_{j=1}^k \mathbf{1}(x \in W_j(h_0))\langle x - c_j, c_j - P_h c_j\rangle$ hence

$$\mathbb{E}_{X\sim\pi}\langle X - P_{h_0} X, P_{h_0} X - P_h P_{h_0} X\rangle = \sum_{j=1}^k \mathbb{E}_{X\sim\pi}\{1(X \in W_j(h_0))\langle X - c_j, c_j - P_h c_j\rangle\}.$$

When h_0 satisfies (143), each term on the right hand side vanishes, either because $\pi(X \in W_j(h_0)) = 0$ or because $c_j = \mathbb{E}_\pi(X|X \in W_j(h_0))$. As a result $\Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(P_{h_0}\pi, h) \leq 0$ for any h. As $D_h^{\overline{\mathcal{H}}}$ is non-negative, we get $D_{h_0}^{\overline{\mathcal{H}}}(\pi \| P_{h_0}\pi) = 0$.

If the support of π contains at least k elements then the unconstrained k-means optimizer $h^\star$ on $\overline{\mathcal{H}}$ satisfies the centroid condition (19), which implies that for $1 \leq j \leq k$ we have $\pi(X \in W_j(h_0)) > 0$ and $c_j = \mathbb{E}_{X\sim\pi}(X|X \in W_j(h_0))$, hence assumption (143) holds and we can use the result established above. It is straightforward to check that (143) holds as well if the support of π contains at most $k-1$ elements, i.e., if π is a mixture of $k-1$ Diracs.

When $\mathcal{H} \subset \mathcal{H}_R$ we have $\langle x - P_{h_0}x, P_{h_0}x - P_hP_{h_0}x\rangle \leq \|x - P_{h_0}x\|_2 \cdot 2R$ hence $\Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(P_{h_0}\pi, h) \leq 4R \cdot \mathbb{E}_{X\sim\pi}\|X - P_{h_0}X\|_2 = \mathcal{R}_{\texttt{k-medians}}(\pi, h_0)$. □

The term $d^{\mathcal{H}}(\pi^\star, \mathfrak{S})$ can also be simplified for clustering when $\mathfrak{S} = \mathfrak{S}^\star(\mathcal{H})$.

Lemma D.7. *Consider $\pi^\star := \sum_{i=1}^k \alpha_i\delta_{c_i}$ where $c_1, \ldots, c_k \in \mathbb{R}^d$, $\alpha \in \mathbb{S}_{k-1}$ and the k-medians (resp k-means) task with a class $\mathcal{H}$. For k-medians we have*

$$\sup_{h'\in\mathcal{H}} \Big(\mathcal{R}_{\texttt{k-medians}}(\pi^\star, h') - \mathcal{R}_{\texttt{k-medians}}(P_h\pi^\star, h')\Big) = \mathcal{R}_{\texttt{k-medians}}(\pi^\star, h), \qquad \forall h \in \mathcal{H}.$$

For k-means and any $h_0 \in \mathcal{H}$ such that (143) *holds with $\pi := \pi^\star$ we have*

$$\sup_{h'\in\mathcal{H}} \Big(\mathcal{R}_{\texttt{k-means}}(\pi^\star, h') - \mathcal{R}_{\texttt{k-means}}(P_{h_0}\pi^\star, h')\Big) = \mathcal{R}_{\texttt{k-means}}(\pi^\star, h_0).$$

If $\mathcal{H} \subseteq \mathcal{H}_R := \{h = (c_1, \ldots, c_k), \|c_l\|_2 \leq R\}$ we further have for each $h \in \mathcal{H}$

$$\mathcal{R}_{\texttt{k-means}}(\pi^\star, h) \leq \sup_{h'\in\mathcal{H}} \Big(\mathcal{R}_{\texttt{k-means}}(\pi^\star, h') - \mathcal{R}_{\texttt{k-means}}(P_h\pi^\star, h')\Big) \leq \mathcal{R}_{\texttt{k-means}}(\pi^\star, h) + 4R\cdot\mathcal{R}_{\texttt{k-medians}}(\pi^\star, h).$$

Proof. Since $\mathcal{R}(P_h\pi^\star, h) = 0$ for both k-means and k-medians we have

$$\begin{aligned}\sup_{h'\in\mathcal{H}} \Big(\mathcal{R}(\pi^\star, h') - \mathcal{R}(P_h\pi^\star, h')\Big) &= \sup_{h'\in\mathcal{H}} \Big(\Delta\mathcal{R}_h(\pi^\star, h') - \Delta\mathcal{R}_h(P_h\pi^\star, h')\Big) + \mathcal{R}(\pi^\star, h)\\ &= D_h^{\mathcal{H}}(\pi^\star, P_h\pi^\star) + \mathcal{R}(\pi^\star, h)\end{aligned}$$

For k-medians, we conclude using that $D_h^{\mathcal{H}}(\pi^\star, P_h\pi^\star) = 0$ by [Gribonval et al., 2021, Lemma 3.2]. Using Lemma D.6 yields the results for k-means. □

D.4 Proof of Theorems 3.1 and Theorem 4.1

With the separated hypothesis class $\mathcal{H}_{\texttt{sep}} := \mathcal{H}_{k,2\varepsilon,R}$ defined in (24) (resp. in (35)), the model set is a separated mixture model, $\mathfrak{S}(\mathcal{H}) := \mathfrak{S}^\star(\mathcal{H}) \subset \mathfrak{S}_k(\mathcal{T})$, where $\mathcal{T} = \mathcal{T}_{\texttt{Dirac}}$ (resp. $\mathfrak{S}(\mathcal{H}) := \mathfrak{S}^{\texttt{ML}}(\mathcal{H}) \subseteq \mathfrak{S}_k(\mathcal{T})$ with $\mathcal{T} = \mathcal{T}_{\texttt{Gauss}}$) as in Definition 6.9., By definition of $\mathcal{H}_{\texttt{sep}}$ (cf. (24) and (35)) the centers c_l associated to any $h \in \mathcal{H}_{\texttt{sep}}$ satisfy $\max_l\|c_l\|_2 \leq R$ (resp. $\|c_k\|_{\boldsymbol{\Sigma}} \leq R$ for GMM) hence the parameter space $\boldsymbol{\Theta}$ is the ball of radius R with respect to the Euclidean norm (resp. the Mahalanobis norm $\|\cdot\|_{\boldsymbol{\Sigma}}$).

The function Φ is defined using the random Fourier feature family $(\mathcal{F}_{\texttt{Dirac}}, \Lambda_{\texttt{Dirac}})$ (resp. $(\mathcal{F}_{\texttt{Gauss}}, \Lambda_{\texttt{Gauss}})$) as in Definition 6.9, with scale factor $s > 0$. By the derivations in Section 6.3.1 the induced average kernel κ is shift-invariant and 1-strongly locally characteristic with respect to $\mathcal{T}$.

We now control the constants $C_{\boldsymbol{\Theta}}, A, B, C$ from Theorem 6.11. By (87) we have

$$\|\pi_0\|_\kappa^{-1} = \begin{cases} C_\Lambda = \big[\mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)} w^2(\omega)\big]^{1/2}, & \text{for Diracs,}\\ \big(1 + 2/s^2\big)^{d/4}, & \text{for Gaussians.}\end{cases} \tag{145}$$

Consider first the Dirac setting. Since $\mathbf{\Theta}$ is the Euclidean ball of radius $R \geq \varepsilon$ (see Definition 6.9), by Lemma A.4 we get that (91) holds with $C_{\mathbf{\Theta}} = 4R$. Moreover, recall that $w(\omega) = (1 + s^2 d^{-1}\|\omega\|^2)$. Then, since $(\varepsilon/s)^2 = 1/(\sigma_k^\star)^2 = 16\log(ek)$, elementary calculations give:

$$\varepsilon \sup_\omega \tfrac{\|\omega\|_2}{w(\omega)} = \varepsilon\sqrt{\frac{d}{4s^2}} = \sqrt{\frac{d}{4(\sigma_k^\star)^2}} = \sqrt{4d\log(ek)}; \qquad \varepsilon^2 \sup_\omega \tfrac{\|\omega\|_2^2}{w(\omega)} = \varepsilon^2 \frac{d}{s^2} = 16d\log(ek);$$
$$\mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)}\|\omega\|_2^2 = s^{-2}d; \qquad \mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)}\|\omega\|_2^4 = s^{-4}(d^2+2d);$$

$$\begin{aligned}
\|\pi_0\|_\kappa^{-2} = A = \mathbb{E}_{\omega\sim\mathcal{N}(0,s^{-2}\mathbf{I}_d)} w^2(\omega) &= 1 + 2s^2 d^{-1}\mathbb{E}_\omega\|\omega\|_2^2 + s^4 d^{-2}\mathbb{E}_\omega\|\omega\|_2^4 = 1 + 2 + (1 + 2/d) \leq 6;\\
B &= 1 + \varepsilon^2\left(\sup_\omega \tfrac{\|\omega\|_2}{w(\omega)}\right)^2 = 1 + 4d\log(ek) \leq 5d\log(ek);\\
C &= 64A\sqrt{2B}C_{\mathbf{\Theta}}\varepsilon^{-1}\left(1 + \varepsilon\sup_\omega \tfrac{\|\omega\|_2}{w(\omega)} + \varepsilon^2 \sup_\omega \tfrac{\|\omega\|_2^2}{w(\omega)}\right)\\
&\lesssim \sqrt{d\log(ek)}(R/\varepsilon)\Big(1 + \sqrt{4d\log(ek)} + 16d\log(ek)\Big) \lesssim (d\log(ek))^{3/2}R/\varepsilon,
\end{aligned}$$

where $\lesssim$ denotes an inequality up to a numerical multiplicative factor. It follows that $\min(12e\log^2(ek), 2B) \lesssim \log(ek)\min(\log(ek), d)$, $kABC \lesssim k(d\log(ek))^{5/2}R/\varepsilon$, and $\log(kABC) \lesssim 1 + \log(kd) + \log(R/\varepsilon)$. As a result there is a numerical constant C' such that (94) holds as soon as

$$m \geq C'\delta^{-2}\log(ek)\min(\log(ek), d)\cdot k\cdot\{kd\cdot[1 + \log(kd) + \log(R/\varepsilon) + \log(1/\delta)] + \log(1/\zeta)\}.$$

Rearranging the terms to put the dominant terms forward, this holds under the assumption (25).

Consider now the GMM setting. As in the Dirac case, since $\mathbf{\Theta}$ is the ball of radius R in the Mahalanobis distance $\|\cdot\|_{\mathbf{\Sigma}}$, by Lemma A.4 we get that (91) holds with $C_{\mathbf{\Theta}} = 4R$. Then, by (90), $\varepsilon^2 = (2 + s^2)/(\sigma_k^\star)^2 = 16(2 + s^2)\log(ek) \asymp s^2\log(ek)$ where $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$, and we have:

$$\begin{aligned}
\|\pi_0\|_\kappa^{-1} = \sqrt{A} &= \big(1 + 2/s^2\big)^{d/4};\\
B &= 1 + \varepsilon^2 \lesssim s^2\log(ek);\\
C &= 64A\sqrt{2B}C_{\mathbf{\Theta}}\varepsilon^{-1}\big(1 + \varepsilon + \varepsilon^2\big) \lesssim R\big(1 + 2/s^2\big)^{d/2}s^2\log(ek).
\end{aligned}$$

It follows that $\min(12e\log^2(ek), 2B) \lesssim \log(ek)\min(\log(ek), s^2)$, $kABC \lesssim Rk(1 + 2/s^2)^d s^4\log^2(ek)$, and $\log(kABC) \lesssim 1 + \log(R) + d/s^2 + \log(ks)$. As a result there is a numerical constant C' such that (94) holds as soon as

$$m \geq C'\delta^{-2}(1+2/s^2)^{d/2}\log(ek)\min(\log(ek), s^2)\cdot k\cdot\{kd\cdot[1+\log(R)+d/s^2+\log(ks)+\log(1/\delta)]+\log(1/\zeta)\}.$$

Rearranging the terms to put the dominant terms forward, this holds under the assumption (40).

We have all ingredients to apply Theorem 6.11 hence, with probability at least $1 - \zeta$ the operator $\mathcal{A}$ induced by Φ satisfies (26).

As a result, under the assumption (25) (resp. (40)) we have all ingredients to apply Theorem 6.11 hence, with probability at least $1 - \zeta$ the operator $\mathcal{A}$ induced by Φ satisfies (26) (resp. (41)).

By (26) (resp. (41)) and the shift-invariance of κ, for any $\theta, \theta' \in \mathbb{R}^d$ such that $\|\theta - \theta'\|_2 \leq \varepsilon$ (resp. $\|\theta - \theta'\|_{\mathbf{\Sigma}} \leq \varepsilon$) we have $\|\mathcal{A}(\pi_\theta) - \mathcal{A}(\pi_{\theta'})\|_2^2 \leq (1 + \delta)\|\pi_\theta - \pi_{\theta'}\|_\kappa^2 = 2\|\pi_0\|_\kappa^2(1 + \delta)(1 - \overline{\kappa}(\theta - \theta'))$ hence

$$\|\mathcal{A}(\pi_\theta) - \mathcal{A}(\pi_{\theta'})\|_2^2 \leq 2\|\pi_0\|_\kappa^2(1 + \delta)\begin{cases} 1 - e^{-\frac{\|\theta-\theta'\|_2^2}{2\varepsilon^2(\sigma_k^\star)^2}} & \text{(for clustering);}\\ 1 - e^{-\frac{\|\theta-\theta'\|_{\mathbf{\Sigma}}^2}{2\varepsilon^2(\sigma_k^\star)^2}} & \text{(for GMM).}\end{cases}$$

As $f : u \mapsto 1 - e^{-\frac{u}{2\varepsilon^2(\sigma_k^\star)^2}}$ is concave we have $f(u) \leq f(0) + uf'(0)$ for each $u \in \mathbb{R}$, hence $\theta \mapsto \mathcal{A}(\pi_\theta)$ is L-Lipschitz with respect to $\|\cdot\|_2$ (resp. $\|\cdot\|_{\mathbf{\Sigma}}$) in $\mathbb{R}^d$ and $\|\cdot\|_2$ in $\mathbb{C}^m$, with $L = \|\pi_0\|_\kappa \sqrt{1+\delta}/(\varepsilon\sigma_k^\star)$. For clustering we have $\mathcal{A}(\pi_\theta) = \mathcal{A}(\delta_\theta) = \Phi(\theta)$ and $\|\pi_0\|_\kappa \leq 1$ (by (145) and the fact that $w \geq 1$) hence the claimed Lipschitz property of Φ.

A second consequence of (26) is the LRIP (11) on $\mathfrak{S}(\mathcal{H}) \subseteq \mathfrak{S}_k(\mathcal{T}))$ with $\eta = 0$ and $C_\mathcal{A} := 8\sqrt{2k/(1-\delta)}\|\mathcal{D}\|_{\Delta\mathcal{L}(\mathcal{H})}$. By Theorem 2.2, since $\hat{h}$ satisfies (27) (resp. (42)), we get

$$\forall h_0 \in \mathcal{H} : \qquad \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) \leq d_{h_0}^{\mathcal{H}}(\pi, \mathfrak{S}(\mathcal{H})) + (2+\nu)C_\mathcal{A}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n))\|_2 + C_\mathcal{A}\nu'.$$

Since $\mathcal{H} \subseteq \overline{\mathcal{H}}$ denoting $\pi^\star := P_{h^\star}\pi$ (resp. $\pi^\star := \pi_{h^\star}$), we have by Lemma D.5, with $d(\pi^\star, \mathcal{H})$ as in (139) (resp. as in (140)):

$$\begin{aligned}
\Delta\mathcal{R}_{h^\star}(\pi, \hat{h}) &= \Delta\mathcal{R}_{h^\star}(\pi, h_0) + \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) = \inf_{h_0 \in \mathcal{H}_{\text{opt}}} \Big\{ \Delta\mathcal{R}_{h^\star}(\pi, h_0) + \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) \Big\} \\
&\leq (2+\nu)C_\mathcal{A}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n))\|_2 + C_\mathcal{A}\nu' + \inf_{h_0 \in \mathcal{H}} \big\{ \Delta\mathcal{R}_{h^\star}(\pi, h_0) + d_{h_0}^{\mathcal{H}}(\pi, \mathfrak{S}(\mathcal{H})) \big\} \\
&\leq (2+\nu)C_\mathcal{A}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 + C_\mathcal{A}\nu' \\
&\quad + \Big[D_{h^\star}^{\overline{\mathcal{H}}}(\pi\|\pi^\star) + (2+\nu)C_\mathcal{A}\|\mathcal{A}(\pi) - \mathcal{A}(\pi^\star)\|_2 \Big] + d(\pi^\star, \mathcal{H}).
\end{aligned}$$

The excess risk divergence term $D_{h^\star}^{\overline{\mathcal{H}}}(\pi\|\pi^\star)$ vanishes for k-medians by [Gribonval et al., 2021, Lemma 3.4]. Since $\overline{\mathcal{H}} = (\mathbb{R}^d)^k$ it also vanishes for k-means by Lemma D.6.

To conclude, we explicit the involved constants. For clustering since $\|\pi_0\|_\kappa^{-1} \leq \sqrt{6}$, by Lemma D.1 we get $\|\mathcal{D}\|_{\Delta\mathcal{L}(\mathcal{H})} \leq 2(2R)^p\|\pi_0\|_\kappa^{-1} \leq \sqrt{24}(2R)^p$. Since $8\sqrt{2}\sqrt{24} = 8\sqrt{48} \leq 8 \cdot 7 = 56$ we obtain:

$$C_\mathcal{A}^{\texttt{clust.}} \leq 56\sqrt{k/(1-\delta)}(2R)^p. \tag{146}$$

For GMM, Lemma D.2 yields $\|\mathcal{D}\|_{\Delta\mathcal{L}(\mathcal{H})} \leq 4R^2\|\pi_0\|_\kappa^{-1}$. By (145), since $8\sqrt{2}\cdot 4 = 32\sqrt{2} \leq 46$ we obtain

$$C_\mathcal{A}^{\texttt{GMM}} \leq 46\sqrt{k/(1-\delta)}R^2\big(1 + 2/s^2\big)^{d/4}. \tag{147}$$

D.5 Proof of Lemma 3.2

Consider $h \in \mathcal{H}$ and $\tau := \sum_{i=1}^k \alpha_i \delta_{P_h c_i}$. Gathering the terms in τ indexed by $i \neq j$ such that $P_h c_i = P_h c_j$, we see that $\tau \in \mathfrak{S}_h^\star$. For k-medians, by Lemma D.7 and the Lipschitz property of Φ,

$$\begin{aligned}
&\sup_{h' \in \mathcal{H}} \Big(\mathcal{R}_{\texttt{k-medians}}(\pi^\star, h') - \mathcal{R}_{\texttt{k-medians}}(\tau, h') \Big) \leq \mathcal{R}_{\texttt{k-medians}}(\pi, h) \\
&\|\mathcal{A}(\pi^\star) - \mathcal{A}(\tau)\|_2 = \left\| \sum_{i=1}^k \alpha_i \big(\Phi(c_i) - \Phi(P_h c_i) \big) \right\|_2 \leq \sum_{i=1}^k \alpha_i \|\Phi(c_i) - \Phi(P_h c_i)\|_2 \leq \sum_{i=1}^k \alpha_i L \|c_i - P_h c_i\|_2 \\
&\qquad\qquad = L \cdot \mathcal{R}_{\texttt{k-medians}}(\pi^\star, h)
\end{aligned}$$

By the definition (29) of $d(\pi^\star, \mathcal{H})$ this implies $d(\pi^\star, \mathcal{H}) \leq \inf_{h \in \mathcal{H}}(1 + (2+\nu)C_\mathcal{A} L)\mathcal{R}_{\texttt{k-medians}}(\pi^\star, h)$. Turning to k-means, since $\mathcal{H} \subseteq \mathcal{H}_{k,2\varepsilon,R} \subset \mathcal{H}_R := \{h = (c_1, \ldots, c_k), \|c_l\|_2 \leq R\}$ by Lemma D.7 we get

$$\sup_{h' \in \mathcal{H}} \Big(\mathcal{R}_{\texttt{k-means}}(\pi^\star, h') - \mathcal{R}_{\texttt{k-means}}(\tau, h') \Big) \leq \mathcal{R}_{\texttt{k-means}}(\pi, h) + 4R \cdot \mathcal{R}_{\texttt{k-medians}}(\pi, h).$$

The rest of the proof is the same as for k-medians. Since $L \leq \sqrt{1+\delta}/s$, $\varepsilon = 4s\sqrt{\log(ek)}$ and $C_\mathcal{A} \leq 56\sqrt{k/(1-\delta)}(2R)^p$ we have $C_\mathcal{A} L \leq 224\sqrt{k\log(ek)(1+\delta)/(1-\delta)}(2R)^p/\varepsilon$.

Table of notations

$x \in \mathcal{Z}$	sample and sample space
$\mathbf{y}$	sketch (2)
Φ	sketching function (2)
$\mathcal{A}$	sketching operator (6)
π, τ	probabilities on sample space
μ, ν	measures on sample space
$\langle \pi, f \rangle$	$\mathbb{E}_{X\sim\pi} f(X)$
$\langle \mu, f \rangle$	$\int f(x) d\mu(x)$
$\mathrm{KL}(\pi \| \pi')$	KL-divergence (33)
$\mathrm{H}(\pi)$	differential entropy
h	hypothesis
$\mathcal{H} \subseteq \overline{\mathcal{H}}$	classes of hypotheses
$\ell(\cdot, h)$	loss function
$\mathcal{R}$, $\Delta\mathcal{R}_h$	risk (1), excess risk (7)
$h^\star = h^\star_\pi$	best hypothesis (1)
R	generic proxy for the risk (3)
$R_{\texttt{clust.}}$, $R_{\texttt{GMM}}$	specific proxies (4); (5)
$\hat{h}$	learned hypothesis (3)
P_h	projection function for comp.-type task (Sec. 3.1)
$\mathcal{L} = \mathcal{L}(\mathcal{H})$	class of loss functions (Th. 2.2)
$\Delta\mathcal{L} = \Delta\mathcal{L}(\mathcal{H})$	class of loss differences (10)
$\kappa(x, x')$	generic psd kernel (46)
$\kappa(\pi, \pi')$	kernel mean embedding (47)
$\|\mu\|_{\mathcal{G}}$	$\sup_{f\in\mathcal{G}} \lvert\langle \mu, f\rangle\rvert$, (9)
$\|\mu\|_{\kappa}$	MMD norm (48)
$\|\cdot\|_{\Delta\mathcal{L}}$	task-driven norm (10)
$\|\cdot\|$	generic norm on $\mathcal{Z}$
$\|\cdot\|_\star$	dual norm (77)
$\|\cdot\|_2$, $\langle x, x'\rangle$	Euclidean norm, inner product
$\|\cdot\|_{\boldsymbol{\Sigma}}$	Mahalanobis norm (36)
$D_h^{\mathcal{H}}(\pi \| \pi')$	excess-risk divergence (8)
$d_h^{\mathcal{H}}(\pi, \mathfrak{S})$	bias term wrt. model (16)
$d_{\mathcal{F}}(\pi, \pi')$	feature-based metric (52)
$d(\pi, \mathcal{H})$	distance to constraint (29), (44)
$d(\mathbf{c} \| \mathbf{c}'), d(\mathbf{c}, \mathbf{c}')$	distance between k-uples (30)
$\mathfrak{S}$	model set (of probabilities)
$\mathfrak{S}_h^{\texttt{CT}}$, $\mathfrak{S}^{\texttt{CT}}(\mathcal{H})$	compression-type model set (20)
$\mathfrak{S}_h^{\texttt{ML}}$, $\mathfrak{S}^{\texttt{ML}}(\mathcal{H})$	max. likelihood model set (34)
$\mathcal{S} = \mathcal{S}_\kappa(\mathfrak{S})$	normalized secant set (50)
$\boldsymbol{\Theta}$, $\boldsymbol{\Theta}_R$	parameter set (55); Def. 6.9
ϱ	metric on $\boldsymbol{\Theta}$ (55)
φ	embedding (55)
$\mathcal{T}$	$(\boldsymbol{\Theta}, \varrho, \varphi)$ parametric model
$\mathfrak{S}_k(\mathcal{T})$	2-separated k-mixtures (55)
$\mathcal{D} = \mathcal{D}_\kappa(\mathcal{T})$	set of dipoles (56)
$\mathcal{M} = \mathcal{M}_\kappa(\mathcal{T})$	set of monopoles (57)
$\overline{\kappa}(\theta, \theta')$	$\mathcal{T}$-normalized kernel (69)
K, $\mathtt{K}$	kernel-related func. (72),(75)
K_σ	Gaussian kernel (89)
$\mathcal{F} = \{\phi_\omega\}_{\omega\in\Omega}$	generic class of features (45)
Λ	probability distribution on feature parameters ω (46)
$w(\omega)$	weights (Def. 6.3)
C_Λ	normalization constant (85)
s	scale factor (37)
$\sigma(s)$, $\sigma_k^\star$	parameters of Gaussian kernel (89) and (90)
$\mathcal{F}'$, $\mathcal{F}''$	classes derived from Fourier feature class $\mathcal{F}$ (Lem. 6.4 and 6.7)
ε, R	separation, domain bound (24)
$\mathcal{H}_{k,\varepsilon,R}$	constrained hypothesis class (24) (35)
$\pi^\star$	projected distribution (Thm. 3.1, Thm. 4.1)
$V_l(h)$, $W_l(h)$	Voronoi cell, Voronoi partition
$\alpha_l(\pi, h)$	Voronoi weights (18)
$C_{\mathcal{A}}$, $C(K)$, $K_{\max}$, $K'_{\max}$, $K''_{\max}$	constants related to kernel K (11)(74)
$L_{\mathcal{F}}$	Lem. 6.4
$C_{\mathcal{F}}$, $C'_{\mathcal{F}}$, $C''_{\mathcal{F}}$	Lem. 6.7
$\|\mathcal{E}\|$	radius of a set of measures (51)
$c_\kappa(t)$	concentration function (49)
$\mathrm{N}(\|\cdot\|, A, \varepsilon)$	covering numbers
$\mathcal{B}$	ball
$[Y]_{k,\mathcal{W}}$	mixture set (104)

References

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In P. Auer and R. Meir, editors, *Learning Theory (Proceedings of 18th Annual Conference on Learning Theory, COLT 2005)*, pages 458–469. Springer, 2005.

P. Ahrendt. The Multivariate Gaussian Probability Distribution. Technical report, IMM, Technical University of Denmark, 2005.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51(11):4022–4032, 2005.

A. Antos, L. Györfi, and A. Gyorgy. Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51(11):4013–4022, 2005.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM symposium on discrete algorithms*, pages 1027–1035, 2007.

P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813, 1998.

M. Belkin and K. Sinha. Toward learning Gaussian mixtures with arbitrary separation. In A. Kalai and M. Mohri, editors, *Proceedings of 23rd Conference On Learning Theory (COLT)*, pages 407–419, 2010.

M. Bojarski, A. Choromanska, K. Choromanski, F. Fagan, C. Gouy-Pailler, A. Morvan, N. Sakr, T. Sarlos, and J. Atif. Structured adaptive and random spinners for fast machine learning computations. In *Artificial Intelligence and Statistics*, pages 1020–1029. PMLR, 2017.

A. Bourrier, R. Gribonval, and P. Perez. Compressive gaussian mixture estimation. In *ICASSP*, pages 6024–6028, Vancouver, Canada, 2013.

K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

A. Chatalic, R. Gribonval, and N. Keriven. Large-scale high-dimensional clustering with fast sketching. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4714–4718, Calgary, Canada, Apr. 2018. IEEE.

K. Choromanski and V. Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *International Conference on Machine Learning*, pages 2502–2510. PMLR, 2016.

P. A. Chou. The distortion of vector quantizers trained on n vectors decreases to the optimum as $\mathcal{O}(1/n)$. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 457. IEEE, 1994.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. Wiley-Interscience, 1991.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

S. Dasgupta and L. J. Schulman. A two-round variant of EM for Gaussian mixtures. *Uncertainty in Artificial Intelligence*, pages 152–159, 2000.

Y. De Castro, F. Gamboa, D. Henrion, and J.-B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1): 621–630, 2016.

J. Duchi. Derivations for linear algebra and optimization. Technical report, Stanford University, 2007.

V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's Inequality. *IEEE Trans. Inf. Theor.*, 49(6):1491–1498, 2003.

A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *Annals of Statistics*, 5(1):88–97, 1977.

A. Fischer. Quantization and clustering with Bregman divergences. *Journal of Multivariate Analysis*, 101(9):2207–2221, 2010.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2012.

M. R. Garey, D. S. Johnson, and H. S. Witsenhausen. The complexity of the generalized Lloyd - Max problem. *IEEE Trans. Inf. Theory*, 28(2):255–256, 1982.

L. Giffon, V. Emiya, H. Kadri, and L. Ralaivola. Quick-means: accelerating inference for k-means by learning fast transforms. *Machine Learning*, pages 1–25, 2021.

S. Graf, H. Luschgy, and G. Pagès. Optimal quantizers for Radon random vectors in a Banach space. *J. Approx. Theory*, 144(1):27–53, 2007.

R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Compressive statistical learning with random feature moments. *Mathematical Statistics and Learning*, 2021.

N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6190–6194, 2016.

N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval. Compressive k-means. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373, 2017.

N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. *Information and Inference*, 7(3):447–508, 2018.

Q. Le, T. Sarlós, and A. Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *Proceedings of the international conference on machine learning (ICML 2013)*, volume 28, pages 244–252. PMLR, 2013.

C. Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7:1716–1746, 2013.

Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1235–1239. IEEE, 2017.

T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, 1994.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *The 8th International Conference on Probability in Banach Spaces*, volume 30, pages 128–134, 1992.

D. Pollard. Quantization and the method of k-means. *IEEE Trans. Information Theory*, 28(2):199–205, 1982a.

D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926, 1982b.

C. Poon, N. Keriven, and G. Peyré. The geometry of off-the-grid compressed sensing. *arXiv:1802.08464*, 2020.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS 2007)*, volume 20. Curran Associates, Inc., 2008.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS 2008)*, volume 21. Curran Associates, Inc., 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11: 1517–1561, 2010.

H. Steinhaus. Sur la division des corps matériels en parties. *British Journal of Mathematical and Statistical Psychology*, Cl. III — Vol. IV(12):801–804, 1956.

Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all. *Applied and Computational Harmonic Analysis*, 45(1):170–205, 2018.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.